# CERTIFYING TACTICAL MISTAKES IN SELF-PLAY REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Self-play scores can conceal decisions that discard a forced win because sampled opponents fail to refute them. We present a certificate-based audit that turns local tactical proofs into executable opponent responses and quantitative guarantees on their value. The audit distinguishes mistakes on the realized deal from winning actions identifiable using the player's observations. Each denial witness supplies a payoff cap for a responding coalition, reducing certified response optimization to optimal stopping under ordinary self-play. A paired martingale-dual score bounds a fixed switching rule's regret within this response class, using exact conditional values or independent successor simulation. We give a sharp coefficient-one residual bound for predictor-greedy rules and an exact-boundary absorption result. In complete games of a sampled proximal policy optimization (PPO) policy, a simulated switching rule has mean certified score $0.544$ points per game, with stopping regret below $0.0985$ points at $95\%$ confidence. Across $64$ enumerable stochastic endgames, exact regret is $0.00109$ and the paired dual bound is $0.01003$. Refutation replay denies first place on all $1{,}357$ audited throws. As a further use of the audit, reusing the certified action sets in training raises sampled retention from $0.781$ to $0.957$ on a fixed Tien Len challenge set over five independent runs.

## 1 INTRODUCTION

A self-play policy can win after discarding a forced win because the sampled opponents fail to refute its action. Average return records the eventual outcome; it does not identify these forgiven tactical errors. We study tactical reliability through preservation of forced wins and ask how much a coordinated opponent can gain by acting on a certified denial. Card games provide an established setting for self-play (Zha et al., 2021; Li et al., 2020; Lu et al., 2022; Charlesworth, 2018). We study these questions in Big Two, Southern Tien Len and Guandan.

A useful audit must respect the information available to each side. A full-deal solver identifies an action that loses despite a winning alternative. A pooled-opponent game certifies winning alternatives by finding one observable strategy that succeeds across all compatible hidden deals. Separately, a post-action denial witness supplies an executable response using the opponents' joint information. These two proof objects establish an information-feasible correction for the focal player and a payoff guarantee for the responding coalition.

The key operation is to replace the adversarial suffix by that payoff guarantee. The coalition follows the ordinary policy until invoking a certified refutation. Its remaining choice is the switching time, governed by ordinary self-play dynamics. This yields two quantities for evaluation: the response's certified gain and its regret relative to the best rule for the specified certificates and eligibility conditions. A paired representation of the classical martingale dual supplies the regret certificate; shared-prefix cancellation gives a sharp coefficient-one residual envelope, and exact tactical boundaries can remove suffix noise through absorption.

We test this construction through matched refutation replay, exact stochastic calibration, and complete-game evaluation with deterministic and sampled policies. Regression and ordinary-policy simulation construct alternative switching rules, while the same learned martingale can certify either. Guandan tests executable certification with partnership information. The pooled winning-action sets also provide direct supervision: we test their reuse in PPO training through certified-action retention on fixed held-out populations.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our main contributions are:

- We separate true-deal throws, observable winning-action certificates, and executable coalition refutations, turning local tactical proofs into a quantitative opponent-response audit.
- We reduce certified response optimization to ordinary-play stopping and bound the regret of a fixed response rule. The paired dual admits a sharp coefficient-one residual envelope for predictor-greedy rules and exact-boundary absorption that cannot increase the paired score.
- We validate certified responses through exact endgame calibration and full-game regret certificates. Paired training experiments additionally test reuse of the audit's winning-action labels as policy supervision.

## 2 Related work

**Solver-based policy audits.** Perfect Information Post-Mortem Analysis (PIPMA) identifies decisions that turn a win into a loss in card-game episodes (Long & Buro, 2011); we use the same win-to-loss error criterion. Paranoid-error filtering in Skat (Furtak, 2013) and solver labels in chess (Haque et al., 2021; LCZero team, 2018; Ruoss et al., 2024) serve related purposes. Our audit tests whether a winning action is identifiable from the player's information and attaches an executable response with a certified payoff value.

**Planning with hidden information.** Best-defence search distinguishes one strategy that works across hidden deals from separate perfect-information plans (Frank & Basin, 1998; Frank et al., 1998). Combined-defender and knowledge-based paranoia search also strengthen the opponents to certify safe decisions (Bethe, 2010; Furtak, 2013; Edelkamp, 2021). Our pooled game adapts this idea to multi-card shedding games while preserving seat capacities and turn rules. Unlike privileged supervision in PerfectDou and Suphx (Yang et al., 2022; Li et al., 2020), its labels use only execution-time information.

**Adversarial evaluation and targeted training.** Adversarial policies expose failures in strong agents (Gleave et al., 2020; Wang et al., 2023; Tseng et al., 2025), and explicit opponents lower-bound exploitability (Lisý & Bowling, 2017; Timbers et al., 2022). Team-game models make coalition information and coordination explicit (von Stengel & Koller, 1997; Celli & Gatti, 2018; MacQueen & Wright, 2023). Our solver supplies the adversarial suffix, so learning chooses only when to invoke it. Reset curricula (Florensa et al., 2017; Salimans & Chen, 2018; Lanier et al., 2026) and policy-gradient coverage analyses (Kakade & Langford, 2002; Agarwal et al., 2021) motivate training on audited states; our audit also supplies the target actions.

**Optimal stopping.** Simulation-based methods learn continuation values under a fixed process (Tsitsiklis & Van Roy, 1999; Longstaff & Schwartz, 2001). Martingale duality bounds the optimal value from approximate functions (Rogers, 2002; Haugh & Kogan, 2004), with nested simulation providing practical estimators (Andersen & Broadie, 2004). Recent work also studies favorable occupancy for stationary discounted stopping (Bhandari & Russo, 2024). We apply these tools to an executable coalition response and express the classical primal–dual gap as a nonnegative paired score, with a coefficient-one residual bound and exact-boundary absorption.

## 3 Setting and audit

Consider an $N$-player game with a fixed deal and a terminal payoff $u_h$ for a focal player $h$. The complete state $s$ contains every hand and the public history. The focal player's information state $I$ contains only its hand and public observations, and its policy is $\pi(a \mid I)$. In the Big Two and TLMN audits, all cards are dealt and the other three players form a coalition whose shared observations determine the full deal. They coordinate using their joint history. Appendix I gives the partnership-information formulation for Guandan.

Let $\Sigma_C$ be the set of strategies measurable with respect to the coalition's joint history; it includes the strategy that uses $\pi$ at every coalition seat. We define the ordinary self-play value $J_{\mathrm{self}}(\pi)$ and the coalition's gain $L_{\mathrm{col}}(\pi)$ by

$$L_{\mathrm{col}}(\pi) = J_{\mathrm{self}}(\pi) - \inf_{\sigma \in \Sigma_C} \mathbb{E}[u_h(\pi, \sigma)], \qquad J_{\mathrm{self}}(\pi) = \mathbb{E}_\pi[u_h].$$

**Mistakes on the realized deal.** Exact paranoid search (Sturtevant & Korf, 2000) treats the other players as a single adversary. At a focal decision, it partitions the actions into proven-winning actions $A^+(s)$, proven-losing actions $A^-(s)$, and—when search has a node limit—unresolved actions. We define the denial and throw indicators as

$$\mathsf{Deny}(s,a) = \mathbf{1}\{a \in A^-(s)\}, \qquad \mathsf{Throw}(s,a) = \mathbf{1}\{A^+(s) \neq \varnothing\}\,\mathsf{Deny}(s,a).$$

A *denial*, $\mathsf{Deny}(s,a) = 1$, means that the opponents can prevent the focal player from finishing first after action $a$. A *throw*, $\mathsf{Throw}(s,a) = 1$, is a denial at a state with at least one winning alternative. We report throw rates conditional on proven-win decisions and list unresolved actions separately.

**Actions certified under hidden information.** Let $\mathfrak{D}(I)$ denote all deals consistent with information state $I$. To obtain a conservative certificate, we replace the opponents' separate hands by their union but preserve each opponent's remaining number of cards. On its turn, an opponent may play any legal combination from this pool that fits its remaining capacity. All pass, lead, and termination rules remain unchanged. Solving this harder game gives $A^{\mathrm{pool}}(I)$, the set of first actions belonging to a winning strategy, and $\mathrm{SW}_{\mathrm{pool}}(I) = \mathbf{1}\{A^{\mathrm{pool}}(I) \neq \varnothing\}$, the indicator that such a strategy exists. Because this certificate depends only on $I$, a player can use it at execution time. We call $\mathsf{Throw}(s,a)\mathrm{SW}_{\mathrm{pool}}(I) = 1$ an *information-chargeable throw*: the chosen action loses on the realized deal even though the player could identify a certified winning action from its observations.

**Win retention.** For any action set $A$, write $\pi(A \mid I) := \sum_{a \in A} \pi(a \mid I)$. For a fixed distribution $\nu$ over certified positions, we measure the probability mass that policy $\pi$ assigns to true-deal and pooled winning actions:

$$\rho_{\mathrm{deal}}(\pi;\nu) = \mathbb{E}_{s\sim\nu}\pi(A^+(s) \mid I), \qquad \rho_{\mathrm{pool}}(\pi;\nu) = \mathbb{E}_{s\sim\nu}\pi(A^{\mathrm{pool}}(I) \mid I).$$

For greedy evaluation, we replace each probability by an indicator for the selected action. With fully resolved true-deal labels, pooled soundness implies $\rho_{\mathrm{pool}} \leq \rho_{\mathrm{deal}}$ on the same positions and under the same behavior mode. Under capped search, pooled proofs can instead be added to $A^+$. The difference between the two measures is the probability assigned to actions that win on the realized deal but lack a pooled certificate.

**Can ordinary opponents exploit the mistake?** For each recorded throw $(s,a)$, we force action $a$ and compare two continuations. In the first, every player follows the audited policy. In the second, the opponents execute the solver's refutation. The *mirror denial probability* $p_{\mathrm{pun}}(s,a)$ is the probability that the focal player fails to finish first in the first continuation. The second continuation checks that the certified refutation can be executed against the focal policy's later decisions.

**Running example: a mistake that self-play rewards**

Consider a Big Two information state $I$ in which the focal player has three cards. Its policy chooses action $a_{\mathrm{throw}}$, leaving two cards. On the realized deal $s$, exact search finds another legal action $a_{\mathrm{safe}} \in A^+(s)$ but proves $a_{\mathrm{throw}} \in A^-(s)$. Thus $\mathsf{Throw}(s, a_{\mathrm{throw}}) = 1$: the policy has discarded a forced win. Nevertheless, sampled mirror opponents may later pass, allowing the focal player to shed its remaining cards and receive, for example, payoff $U = 3$. Outcome-only self-play would therefore reward a decision that an exact opponent can refute.
The full-deal result alone does not show that the focal player could identify the better action from $I$. We therefore solve the pooled-opponent game constructed from $I$. If $a_{\mathrm{safe}} \in A^{\mathrm{pool}}(I)$, the same observable strategy wins for every hidden deal consistent with $I$. After $a_{\mathrm{throw}}$, the solver also supplies an executable refutation that guarantees the focal payoff is at most $B_t = -1$. On the trajectory above, switching to this refutation exposes a gain of at least $U - B_t = 4$ for the coalition. This example will illustrate the remaining two steps: learning when to switch and using the certified action for policy supervision.

## 4 Certified opponent responses

An executable denial witness supplies a bound on the focal player's payoff after a switch. We use that bound as a boundary cost: the coalition follows ordinary play until selecting a certified refutation. Optimizing this certified response value therefore requires choosing the switching time. We evaluate both the gain guaranteed by a chosen rule and its remaining regret in the resulting stopping class.

Let $\mathcal{F}_t$ denote the information available to the coalition at time $t$ during ordinary all-$\pi$ play. It includes the focal player's observed action and any fresh randomness used by the switching rule. The game ends by a finite horizon $H$ and has bounded terminal focal-player payoff $U$. Define the boundary process $B = (B_t)_{t=0}^{H}$ as follows. At an eligible denial, $B_t$ is a certified upper bound on the focal player's *total* payoff if the coalition switches to the refutation; at an ineligible time $t < H$, $B_t = +\infty$; and at termination, $B_H = U$. All finite caps are uniformly bounded, and both eligibility and $B_t$ are measurable with respect to $\mathcal{F}_t$. For example, every denial before the first finish in Big Two v2 has the valid cap $B_t = -1$.

An adapted stopping rule $\tau$ selects an eligible switching time using only the history available to the coalition. The coalition follows $\pi$ before $\tau$ and executes the refutation afterward; $\tau = H$ means that it never switches. Define the cap-valued payoff

$$C_\tau = \begin{cases} B_\tau, & \tau < H, \\ U, & \tau = H. \end{cases} \qquad F(\tau) = \mathbb{E}_\pi[U - C_\tau].$$

Because ordinary and switching play are identical before $\tau$, we can evaluate this response on an ordinary trajectory. The certificate guarantees $L_{\mathrm{col}}(\pi) \geq F(\tau)$. Hence, maximizing the certified coalition gain is equivalent to minimizing the cap-valued payoff: $F^* = \mathbb{E}U - \inf_\tau \mathbb{E}C_\tau$.

**The certified stopping objective.** The stopping problem uses only transitions generated by ordinary self-play:

$$v_H^* = U, \qquad q_t^* = \mathbb{E}_\pi[v_{t+1}^* \mid \mathcal{F}_t], \qquad v_t^* = \min\{B_t, q_t^*\}. \tag{1}$$

For $t = 0, \ldots, H-1$, $q_t^*$ is the conditional cost of continuing ordinary play for one more step while retaining the option to switch later, and $v_t^*$ is the optimal stopped cost. The coalition switches at the first time $B_t \leq q_t^*$, and the optimal certified gain is $F^* = \mathbb{E}[U - v_0^*]$. The certificate removes the need to model the adversarial suffix; one continuation function captures all future switching opportunities. Equation (1) is the standard optimal-stopping recursion (Tsitsiklis & Van Roy, 1999; Longstaff & Schwartz, 2001). We next use its martingale dual (Rogers, 2002; Haugh & Kogan, 2004) to evaluate an approximate continuation predictor without knowing $q^*$.

**Theorem 4.1** (Certified response interval)**.** *Fix a bounded adapted predictor $q = (q_t)_{t<H}$. Set $v_t = \min\{B_t, q_t\}$ for $t < H$ and $v_H = U$. Let $\tau_q$ be the first eligible time with $B_t \leq q_t$, or $H$ if there is none. Let $\mathcal{T}$ contain all eligible times and $H$, and define*

$$\delta_t = q_t - \mathbb{E}_\pi[v_{t+1} \mid \mathcal{F}_t], \qquad M_t = \sum_{s<t} \big(v_{s+1} - \mathbb{E}_\pi[v_{s+1} \mid \mathcal{F}_s]\big),$$

*and $Z_q = C_{\tau_q} - M_{\tau_q} - \min_{t\in\mathcal{T}}(B_t - M_t)$. Then $Z_q \geq 0$ on every ordinary path and*

$$0 \leq F^* - F(\tau_q) \leq \mathcal{D}(q) := \mathbb{E}_\pi Z_q \leq \mathbb{E}_\pi \sum_{t=0}^{H-1} |\delta_t|. \tag{2}$$

*The coefficient 1 is sharp, and $Z_{q^*} = 0$ pathwise.*

The audit now has two quantitative outputs: $F(\tau_q)$ lower-bounds the gain of an executable coalition, and $\mathcal{D}(q)$ bounds the rule's regret within the certificate-defined stopping class. Their sum bounds $F^*$. When every invoked cap is attained, $F(\tau)$ is that coalition's actual gain. The same martingale can also certify a separately fixed switching rule, including one constructed by ordinary-policy forecasting. Appendix A.5 gives the independent-successor version. The residual bound uses exact conditional means; independent-successor evaluation certifies regret through the expected paired score and its complete-game confidence bound.

**Why the residual coefficient is one.** Candidate stops and $\tau_q$ share a path prefix, so its residuals cancel. Each remaining residual enters at most once, yielding the sharp coefficient in (2); the one-sided refinement and proof are in Appendix A.3. Subtracting the stopped martingale centers the classical primal–dual gap without changing its expectation. We evaluate conditional means exactly or by independent successor simulation, under the conditions in Appendix A.5.

**An exact last-card rule.** Suppose that a denial in Big Two v2 leaves the focal player with one card. Every certified refutation then gives that player payoff $-1$, while ordinary continuation gives at least $-1$. Switching immediately is therefore optimal at this state, and its optimal stopped cost is $-1$. Let

$g$ be any rule that switches only at such denials, let $T_g$ be its first trigger time ($T_g = \infty$ if it never triggers), and let $U^g$ be the focal player's terminal payoff against the induced coalition strategy $\sigma_g$. Finally, let $W$ be the event that the focal player finishes first under ordinary play. Then

$$U - U^g = \mathbf{1}\{T_g < \infty\}(U + 1) = \mathbf{1}\{T_g < \infty, W\}(U + 1) \geq 0, \tag{3}$$

Thus, every additional last-card trigger can only increase the gain on each trajectory. The corresponding reduction in first-place probability is $\mathbb{P}(T_g < \infty, W)$. We use this exact rule both as a direct audit and as an anchor when learning earlier switches. Appendix A contains the proofs and support bound.

The exact boundary also shortens certification. We stop the dual process at the first last-card denial and assign terminal value $-1$. Because every later cap and ordinary payoff is at least $-1$, this absorption preserves both the optimum and the regret of any anchored rule. It removes later candidate times from the dual minimum without changing the shared prefix, so the dual score cannot increase on any path, including under coupled successor simulation (Proposition A.3). In effect, the exact boundary removes irrelevant suffix noise from the certificate.

**An exact rule for deterministic policies.** Suppose ordinary play is deterministic given the coalition's history and every eligible cap equals the same value $b$. Let $\mathcal{A}$ be the event that an eligible boundary occurs; recall that $(x)_+ = \max\{x, 0\}$. From the first eligible state, an ordinary rollout determines the terminal payoff $U$. Switching if and only if $U > b$ attains

$$F^* = \mathbb{E}_\pi[\mathbf{1}\{\mathcal{A}\}(U - b)_+], \tag{4}$$

Since every switch has cap-valued gain $U - b$, this rule is optimal on every path within the prescribed stopping class. Appendix A.7 gives the admissibility argument. The released DMC policy satisfies these conditions with $b = -1$, providing an exact full-game benchmark for learned continuation predictors.

**Executing the audit.** Algorithm 1 makes the certificate-to-response construction explicit. Freeze the ordinary policy, the switching rule $\rho$, the bounded predictor $q$, and the eligibility/search protocol before evaluation. On the eligible-history subsequence, let $v_{s+1}^{(j)}$ be the next eligible value, or terminal value, in auxiliary continuation $j$ drawn from the current coalition history. The independent-successor correction is

$$\widehat{m}_s = \frac{1}{k}\sum_{j=1}^{k} v_{s+1}^{(j)}, \qquad \widehat{M}_t = \sum_{s<t}(v_{s+1} - \widehat{m}_s).$$

Each complete ordinary game supplies the two scores

$$X_\rho = U - C_\rho, \qquad Z_{\rho,q,k} = C_\rho - \widehat{M}_\rho - min_{t\in\mathcal{T}}(B_t - \widehat{M}_t).$$

The first estimates the certified gain of the deployed response. The second gives an expected upper bound on its stopping regret. Conditional independence and the reveal order in Appendix A.5 justify the sampled martingale. For a boundary-anchored switching rule, exact-boundary absorption uses the boundary as the terminal value for the dual while retaining the complete-game payoff $U$ for $X_\rho$.

A deployed response follows $\pi$ until the same gate triggers and then executes the stored refutation. The evaluation therefore attaches its lower bound to a concrete coalition strategy. Every assigned game contributes to the score averages, including negative cap scores and games without an eligible boundary. The gain limits use Appendix B; the independent-successor regret limits use Appendix F.4. Rule construction and certification can share a predictor, while the gate's decision simulations and the dual's auxiliary continuations use separate random streams.

**Constructing switching rules.** For coalition information $I_C$, continuation regression uses the next eligible boundary $B'$ and information $I'_C$ to fit $\min\{B', q_{\text{old}}(I'_C)\}$, where $q_{\text{old}}$ is the preceding fitted predictor; its terminal target is $U$. Outcome regression and independently simulated ordinary suffixes provide alternative forecasts for a switching gate. We select each rule using separate validation games, freeze it, and evaluate its signed score on complete test games. The fitting targets, last-card anchors and margins are specified in Appendix F. The finite-support calculation in Appendix B gives simultaneous lower confidence bounds for a fixed family of rules.

**Algorithm 1** Certified response evaluation on ordinary self-play

**Require:** frozen policy $\pi$, admissible gate $\rho$, bounded predictor $q$, certified-boundary protocol, $n$ independent evaluation games, inner budget $k$

1: **for** each of the $n$ assigned evaluation games **do**
2: Complete ordinary all-$\pi$ play; retain the coalition histories and terminal payoff $U$.
3: At queried histories, record proved eligible caps and their executable witnesses; retain unresolved searches as unknown.
4: Apply $\rho$ using only the available history and its own independent decision draws; record its first switch, or termination.
5: Compute $X_\rho = U - C_\rho$, including zero for a game with no switch.
6: At each dual transition, draw $k$ auxiliary ordinary continuations independently of primary play and gate draws; estimate the next value by $\widehat{m}_s$.
7: Form $\widehat{M}$ and $Z_{\rho,q,k}$; apply exact-boundary absorption when $\rho$ stops by that boundary and Proposition A.3's conditions hold.
8: **end for**
9: Report complete-game means and support-based confidence limits using each endpoint's declared confidence family.

**Running example continued: when should the coalition switch?**

At the post-action denial above, stopping has certified cost $B_t = -1$, while $q_t$ predicts the focal payoff obtained by continuing ordinary play while retaining the option to switch later. Because the coalition minimizes the focal payoff, it switches now when $B_t \le q_t$. For instance, $q_t = 0.2$ triggers the refutation, whereas $q_t = -2$ favors waiting for a better future opportunity. The resulting rule is assessed by its certified gain and its stopping regret.

## 5 Reusing certified action sets

To reuse the audit as policy supervision, form the pool $\mathcal{C} = \{(I_j, A^{\mathrm{pool}}(I_j))\}_j$ of information states and their pooled winning-action sets. We reuse these pairs as training targets in two ways. First, a fraction $\zeta \in [0, 1]$ of episodes starts from a certified state; all other episodes start from ordinary deals. When resetting to a certified state, we resample the hidden cards conditional on the stored observations. The learner therefore sees different hidden deals rather than repeatedly seeing the deal from which the state was collected. The pooled certificate remains valid for every admissible resampling. Second, for policy parameters $\theta$, we add an auxiliary loss that moves the probability assigned by $\pi_\theta$ toward the certified action set:

$$\ell_{\mathrm{cert}}(\theta; I) = -\log \sum_{a \in A^{\mathrm{pool}}(I)} \pi_\theta(a \mid I).$$

Big Two uses this set-mass loss directly. TLMN uses the cross-entropy surrogate $\ell_{\mathrm{CE}} = -\mathbb{E}_{a \sim \mathrm{Unif}(A^{\mathrm{pool}}(I))} \log \pi_\theta(a \mid I)$, where $\mathrm{Unif}(A)$ denotes the uniform distribution over a finite action set $A$. We estimate this loss with one sampled certified action. Both objectives encourage certified actions; the surrogate additionally encourages a uniform distribution within the certified set (Appendix G). At the reset decision, we add the relevant auxiliary loss with nonnegative weight $\lambda$ to the PPO loss. We construct the pool and its held-out split before training the certified variant. Appendix C describes the dataset-specific grouping rules.

Resets bring audited positions into the learner's experience, while the auxiliary loss identifies certified actions even when mirror opponents forgive a throw. We evaluate this joint intervention through certified-action retention at fixed training budgets.

## 6 Experiments

We evaluate the audit by executing certified refutations, measuring the value and regret of switching responses, and reusing the winning-action labels for targeted policy supervision. Big Two v1 uses a released proximal policy optimization (PPO) implementation (Charlesworth, 2018); Big Two v2 uses the deep Monte Carlo (DMC) policy from big2-rl (Chow & Cheng, 2022). Southern Tien Len (TLMN) is a related shedding game with multi-card runs and chop transfers. Training comparisons use five paired runs at the same decision budget. Throughout the tables, SE and CI denote

| Game | Policy | Throws replayed | Mirror denial | Solver denial |
|---|---|---|---|---|
| Big Two v1 | Released PPO | 160 | 73.1% | 100% |
| Big Two v2 | Released DMC | 579 | 75.8% | 100% |
| Big Two v2 | PPO from scratch | 180 | 65.6% | 100% |
| Big Two v2 | PPO clone fine-tune | 206 | 68.0% | 100% |
| TLMN | Frozen self-play reference | 232 | 76.3% | 100% |

Table 1: Executing the same recorded throws against ordinary mirror opponents and a certified refutation. All 1,357 refutations deny first place, with zero solver fallbacks. Each row uses the policy's stated behavior mode.

standard error and confidence interval. Appendix C gives the complete experimental protocol, and Appendix D walks through one recorded position per game.

### 6.1 Do ordinary opponents punish certified mistakes?

We first measure throws along trajectories generated by each audited policy. In the exact-search regime, released DMC throws $579/6{,}213 = 9.3\%$ of decisions at which a win is proven to be available. The released PPO policy throws at $2.2\%$ of such decisions; $6.0\%$ of its candidate decisions remain unresolved under the node cap. Our TLMN reference policy throws $232/8{,}917 = 2.6\%$, with no unresolved decisions in the audited slice at a $250{,}000$-node cap. These rates condition on the policy reaching a state with a proven winning action.

Table 1 checks whether the certified punishment can be executed. For each throw, the mirror and solver continuations start from the same state and force the same mistaken action. Copies of the audited policy still let the focal player win in $23.7$–$34.4\%$ of these cases. The solver refutation denies first place in every case. This gap shows why ordinary self-play outcomes alone are insufficient for tactical evaluation.

**Extension to partnership play.** Guandan tests certificates under partnership information: the responding players share their two hands and public history, while the opposing partnership's hand allocation remains hidden. A $20{,}000$-round audit certifies $644$ denials from $1{,}428$ eligible queries. Every certified refutation executes within its payoff cap, and exhaustive physical-action replay verifies all $644$ witnesses. Appendix I gives the full protocol, payoff results, and strength evaluation against DanZero.

### 6.2 Certified gain and stopping regret

We first calibrate the regret certificate where exact values are available, then apply it to complete games. Regression and ordinary-policy forecasting construct alternative switching rules whose certified scores are measured on the same test cohorts.

**Exact calibration in stochastic endgames.** Across $64$ exactly enumerable Big Two v2 endgames, all $1{,}280$ fitted predictors satisfy the ordering in Theorem 4.1. At the largest training budget, the true regret is $0.00109$, the paired dual bound is $0.01003$, and the absolute residual bound is $0.24761$. Appendix H.2 reports the complete learning curves and enumeration protocol.

**A full-game optimum available for validation.** Under the frozen deterministic DMC policy, an ordinary rollout from the coalition's current history determines the terminal payoff. The first-denial oracle in Equation (4) therefore attains the constant-cap stopping optimum on each path. Table 2 compares this executable reference with the learned rules on $60{,}000$ held-out games. Its simultaneous intervals cover both certified gain and the paired regret relative to the oracle. This benchmark evaluates regret directly from complete-game scores, complementing the dual upper bounds in stochastic play. All $512$ independently selected witness executions respect the payoff cap; Appendix F gives the fitting protocol, exact forecast checks and residual diagnostics.

**Complete games with a sampled policy.** We freeze a PPO policy at its native sampling temperature. Three development cohorts each contain $48{,}000$ training and $12{,}000$ validation games. All learned methods use the same sufficient features, $64$ trees, and last-card anchor. A more expensive forecasting rule instead averages $8$ or $32$ independent policy rollouts. We freeze every rule before

| Endpoint | Mean | SE | Simultaneous exact 95% interval |
|---|---|---|---|
| Deterministic oracle | 0.51205 | 0.00824 | [0.47954, 0.54642] |
| Frozen anchored predictor | 0.40180 | 0.00780 | [0.36693, 0.43755] |
| Frozen Monte Carlo predictor | 0.41103 | 0.00789 | [0.37556, 0.44738] |
| Last-card denial | 0.31042 | 0.00670 | [0.28453, 0.33822] |
| Oracle minus anchored | 0.11025 | 0.00318 | [0.09688, 0.12561] |
| Oracle minus Monte Carlo | 0.10102 | 0.00297 | [0.08845, 0.11558] |
| Oracle minus last-card | 0.20163 | 0.00501 | [0.18224, 0.22299] |

Table 2: Constant-cap scores and regrets on the original independent DMC cohort. All three regrets are nonnegative gamewise and bounded by 14.

| Switching rule | Mean bound | Game SE | Simultaneous 95% lower |
|---|---|---|---|
| First last-card denial | 0.13557 | 0.00427 | 0.11778 |
| Anchored / recursive continuation | 0.34932 | 0.00748 | 0.31029 |
| Outcome regression | 0.34623 | 0.00761 | 0.30629 |
| Cash-flow regression | 0.35176 | 0.00736 | 0.31384 |
| Ordinary suffix forecast, 8 branches | 0.54183 | 0.00838 | 0.50149 |
| Ordinary suffix forecast, 32 branches | 0.54430 | 0.00838 | 0.50425 |

Table 3: Certified scores on complete sampled-policy games, in points per game. Learned rows average three independently developed frozen gates within each game. Recursive continuation selects its first anchored fit in every development cohort. Exact finite-support lower bounds cover all seven endpoints and six fixed paired contrasts. Appendix F.3 separates development variation, all stages and cost.

collecting a common 60,000-game test set. All games finish, and all 211,338 candidate decisions are solved exactly.

Anchored switching improves over the last-card baseline by 0.21375 points per game, with a simultaneous lower bound of 0.18035. The 32-branch forecast increases the certified score by 0.19498 [0.18693, 0.20304] over the anchored rule. Forecast rollouts use fresh randomness independent of the realized continuation. Every refutation executed on the first 512 eligible test games respects the cap. Appendix F.3 gives runtimes and complete diagnostics.

**Certifying regret in complete stochastic games.** On a fixed prefix of 32,768 test games, we pair the learned and forecast-based rules with a martingale built from a frozen predictor. Independent successor simulation estimates the next eligible value, and the exact last-card boundary removes the irrelevant suffix. Let $Z_k$ denote the resulting paired score when each conditional value uses $k$ independently simulated successors. Table 4 bounds the 32-branch forecast's remaining regret below 0.10 points with simultaneous 95% coverage, without path enumeration. The additional dual evaluation takes 396.78 seconds; Appendix F.4 gives the cost breakdown, construction, and checks.

On this cohort at $k = 32$, pairing reduces the measured variance of the primal–dual gap estimate by a factor of 5.0 relative to omitting the stopped-martingale centering term (Appendix F.4).

For the 32-branch forecast at $k = 32$, the regret limit comprises a mean paired score of 0.012469, a variance term of 0.002113, and a support term of 0.083845 (Appendix F.4). The last term accounts for most of the confidence margin and depends on the number of independent complete games. Increasing the inner budget affects the mean and variance terms; collecting more complete games also reduces the support term. Thus the reported certificate separates switching performance from the sampling effort needed to establish its regret.

### 6.3 Certified-action retention after training

For Big Two, we train on 5,266 certified positions collected from natural self-play and evaluate on 5,289 held-out positions. For TLMN, we train on 373 positions and evaluate on a curated challenge set of 195 positions. The training and challenge sets have disjoint focal hands and information keys. We select challenge positions where the reference policy fails under the original bounded-search template labels, then certify the final action sets with unbounded-depth search. On the same saved positions, we evaluate both true-deal retention $\rho_{\mathrm{deal}}$ and pooled retention $\rho_{\mathrm{pool}}$. We measure

| Paired switch | Gain | Mean $Z_8$ | Mean $Z_{32}$ | 95% regret upper, $k = 32$ |
|---|---|---|---|---|
| Anchored continuation | 0.34628 | 0.21015 | 0.20950 | 0.30965 |
| Ordinary forecast, 32 branches | 0.54565 | 0.01375 | 0.01247 | **0.09843** |

Table 4: One learned martingale certifies two fixed rules on the same complete sampled-policy games. The forecast uses the general-rule score $Z_{\rho,q,k}$ from Appendix A.5. The dual uses $k = 8/32$ fresh successors, independent of the forecast gate's own 32 suffixes. Simultaneous coverage includes both rules and both inner budgets; the $k = 8$ upper bounds are 0.31037 and 0.09997. Mean dual scores include inner-simulation noise.

| Population | Mode | Metric | Control | Certified | Paired change [95% CI] |
|---|---|---|---|---|---|
| Big Two natural | greedy | $\rho_{\text{deal}}$ | 0.9771 | 0.9778 | $+0.0007$ $[-0.0007, 0.0021]$ |
| | greedy | $\rho_{\text{pool}}$ | 0.9348 | 0.9372 | $+0.0025$ $[0.0004, 0.0045]$ |
| | sampled | $\rho_{\text{deal}}$ | 0.9763 | 0.9767 | $+0.0004$ $[-0.0003, 0.0011]$ |
| | sampled | $\rho_{\text{pool}}$ | 0.9341 | 0.9361 | $+0.0020$ $[0.0006, 0.0033]$ |
| TLMN challenge | greedy | $\rho_{\text{deal}}$ | 0.9733 | 0.9938 | $+0.0205$ $[0.0127, 0.0283]$ |
| | greedy | $\rho_{\text{pool}}$ | 0.8933 | 0.9723 | $+0.0790$ $[0.0558, 0.1022]$ |
| | sampled | $\rho_{\text{deal}}$ | 0.9615 | 0.9903 | $+0.0288$ $[0.0251, 0.0325]$ |
| | sampled | $\rho_{\text{pool}}$ | 0.7807 | **0.9567** | $+\mathbf{0.1760}$ $[0.1565, 0.1954]$ |

Table 5: Five-run final-checkpoint retention at matched training decisions. Each contrast fixes the population, action mode and winning-action definition. Intervals are paired Student-$t$ intervals over training runs, conditional on the benchmark. The sampled columns estimate first-action probability mass.

greedy actions and policy probability estimated from 128 sampled actions per position, using shared randomness for paired checkpoints.

The largest effect is on actions certified under hidden information (Table 5). On the TLMN challenge set, sampled pooled retention increases from 0.7807 to 0.9567, with a two-way bootstrap interval of $[0.1403, 0.2128]$ for the gain. True-deal retention also increases from 0.9615 to 0.9903. On a separate natural set of 2,422 positions, pooled retention is already 0.9899 and does not change significantly. The improvement is therefore concentrated on the intended tactical challenge set.

On Big Two, standard RL already reaches 0.9771 greedy true-deal retention, compared with 0.8313 for DMC. When fine-tuning a DMC behavior clone, certified training reaches 0.9821, a gain of 0.0023 [0.0016, 0.0029] over standard RL. Appendix G reports the complete learning curves.

Appendix G reports training cost, competitive payoffs, and uncertainty.

**Validation and ablations.** On an exactly enumerable Big Two subset, the pooled solver certifies all 328 positions that win under every hidden split. In MiniShed, certified resets and supervision improve retention by 0.0240 over resets alone. Appendix H gives the controls, coverage analysis, and exhaustive checks.

## 7 Conclusion

Local tactical certificates reveal self-play mistakes that sampled opponents can forgive. Our audit distinguishes realized-deal throws from winning actions identifiable under hidden information, and turns denial witnesses into executable opponent responses with payoff caps. The caps reduce certified response selection to optimal stopping under ordinary play. A paired martingale score bounds a fixed rule's regret in this response class, with a sharp coefficient-one residual bound and exact last-card absorption. All 1,357 replayed refutations deny first place. In complete games with a sampled PPO policy, a 32-branch switching rule has mean certified score 0.544 points per game and a 95% stopping-regret upper bound of 0.0985. Reusing certified actions in training raises sampled TLMN challenge-set retention from 0.781 to 0.957.

Our guarantees cover solver-certified states and the specified coalition information and switching class. Scaling certification to larger or less observable games, tightening payoff caps, and using adaptive responses in training are promising next steps.

### Statement of Use of AI

The research questions, the main ideas, the methodology, and the experimental protocols of this paper were chosen by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions on claims and conclusions. Under the authors' direction, we used generative AI tools as assistants for the following tasks with required disclosure: stress-testing mathematical claims and proof arguments, and implementing and running the experiments. In the theoretical component, generative AI was used in part as a "theory attacker": it assisted in implementing computational searches and large-scale simulations over extensive collections of generated examples in order to identify potential counterexamples, edge cases, or weaknesses in candidate theorem statements and proof arguments. These searches were used to help the authors test and refine the theoretical claims. The failure of such searches to identify a counterexample does not constitute a proof, and no theorem in this paper is considered established on the basis of AI-assisted simulation, search, or testing.

The authors independently verified all theorem statements and proofs, citations, experimental specifications, and AI-assisted text against their research artifacts. No experimental data or results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature discovery and for drafting and editing the manuscript, including checks of notation and of the consistency of reported numbers. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

### Ethics statement

The coordinated strategies studied here are evaluation opponents for auditing game-playing policies. Experiments use simulated games and released research checkpoints.

### Reproducibility statement

Sections 3–5 specify the tactical audit, the certified opponent response and its evaluation (Algorithm 1), and the reuse of certified actions in training. Appendix A gives complete proofs; Appendix B gives the finite-sample confidence construction. Appendix C specifies game rules, information states, model architectures, search and training budgets, reset distributions, action modes, payoffs and evaluation cohorts, with recorded-game examples in Appendix D. Section 6 summarizes the experiments. Appendices E–I give the full-game audits, frozen stopping protocols, training controls, small-game calibration and partnership evaluation, together with the uncertainty calculations and executable checks supporting the reported results. The supplementary material provides the source code for the audits, the certified-response and stopping studies, the training configurations, the Guandan audit and the small-game study.

## References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Leif Andersen and Mark Broadie. Primal-dual simulation algorithm for pricing multidimensional american options. *Management Science*, 50(9):1222–1234, 2004. doi: 10.1287/mnsc.1040.0258.

Paul M. Bethe. DTAC: A method for planning to claim in Bridge. Master's thesis, New York University, 2010. URL `https://cs.nyu.edu/media/publications/bethe_paul.pdf`.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024. doi: 10.1287/opre.2021.0014.

Andrea Celli and Nicola Gatti. Computational results for extensive-form adversarial team games. In *AAAI*, 2018.

Henry Charlesworth. Application of self-play reinforcement learning to a four-player game of imperfect information, 2018. arXiv preprint.

Jasper Chow and Johnny Cheng. Deep reinforcement learning for Big Two. Software, version 1.0.0, Zenodo, doi:10.5281/zenodo.7811506, https://github.com/johnnyhoichuen/big2-rl, 2022.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Stefan Edelkamp. Knowledge-based paranoia search in trick-taking, 2021. arXiv preprint.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Proceedings of the 1st Conference on Robot Learning*, pp. 482–495, 2017.

Ian Frank and David Basin. Search in games with incomplete information: A case study using Bridge card play. *Artificial Intelligence*, 100(1–2), 1998.

Ian Frank, David Basin, and Hitoshi Matsubara. Finding optimal strategies for imperfect information games. In *AAAI*, 1998.

Timothy M. Furtak. *Symmetries and Search in Trick-Taking Card Games*. PhD thesis, University of Alberta, 2013.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *ICLR*, 2020.

Rejwana Haque, Ting Han Wei, and Martin Müller. On the road to perfection? Evaluating Leela Chess Zero against endgame tablebases. In *Advances in Computer Games (ACG)*, 2021.

Martin B. Haugh and Leonid Kogan. Pricing american options: A duality approach. *Operations Research*, 52(2):258–270, 2004. doi: 10.1287/opre.1030.0070.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

J. B. Lanier, Nathan Monette, Pierre Baldi, and Roy Fox. Data-augmented game starts for accelerating self-play exploration in imperfect-information games. *arXiv preprint arXiv:2605.14379*, 2026.

LCZero team. TB rescoring. lczero.org development blog, https://lczero.org/blog/2018/09/tb-rescoring/, 2018.

Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering Mahjong with deep reinforcement learning, 2020. arXiv preprint.

Viliam Lisý and Michael Bowling. Equilibrium approximation quality of current no-limit poker bots. In *AAAI Workshop on Computer Poker and Imperfect Information Games*, 2017.

Jeffrey Long and Michael Buro. Real-time opponent modelling in trick-taking card games. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 617–622, 2011. URL https://www.ijcai.org/Proceedings/11/Papers/110.pdf.

Francis A. Longstaff and Eduardo S. Schwartz. Valuing american options by simulation: A simple least-squares approach. *The Review of Financial Studies*, 14(1):113–147, 2001. doi: 10.1093/rfs/14.1.113.

Yudong Lu, Jian Zhao, Youpeng Zhao, Wengang Zhou, and Houqiang Li. DanZero: Mastering GuanDan game with reinforcement learning, 2022. arXiv preprint.

Revan MacQueen and James R. Wright. Guarantees for self-play in multiplayer games via polymatrix decomposability. In *NeurIPS*, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

L. C. G. Rogers. Monte carlo valuation of american options. *Mathematical Finance*, 12(3):271–286, 2002. doi: 10.1111/1467-9965.02010.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, Cannada A. Lewis, Joel Veness, and Tim Genewein. Amortized planning with large-scale transformers: A case study on chess. In *NeurIPS*, 2024.

Tim Salimans and Richard Chen. Learning Montezuma's Revenge from a single demonstration. *arXiv preprint arXiv:1812.03381*, 2018.

Nathan R. Sturtevant and Richard E. Korf. On pruning techniques for multi-player games. In *AAAI*, 2000.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *IJCAI*, 2022.

Tom Tseng, Euan McLean, Kellin Pelrine, Tony T. Wang, and Adam Gleave. Can Go AIs be adversarially robust? In *AAAI*, 2025.

John N. Tsitsiklis and Benjamin Van Roy. Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, 1999.

Bernhard von Stengel and Daphne Koller. Team-maxmin equilibria. *Games and Economic Behavior*, 21, 1997.

Tony T. Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell. Adversarial policies beat superhuman Go AIs. In *ICML*, 2023.

Guan Yang, Minghuan Liu, Weijun Hong, Weinan Zhang, Fei Fang, Guangjun Zeng, and Yue Lin. PerfectDou: Dominating DouDizhu with perfect information distillation. In *NeurIPS*, 2022.

Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, and Ji Liu. DouZero: Mastering DouDizhu with self-play deep reinforcement learning. In *ICML*, 2021.

# A Proofs

## A.1 Pooled soundness

**Proposition A.1** (Information-feasible winning actions)**.** *If* $\mathrm{SW}_{\mathrm{pool}}(I) = 1$*, there is a single strategy from* $I$ *that forces a first-place finish on every deal in* $\mathfrak{D}(I)$*. Every* $a \in A^{\mathrm{pool}}(I)$ *starts such a strategy, and* $A^{\mathrm{pool}}(I) \subseteq A^{+}(s)$ *for each fully solved state* $s$ *consistent with* $I$*.*

*Proof.* Fix a deal $d \in \mathfrak{D}(I)$ and a joint opponent strategy on that deal. Map each reachable real position to the pooled position with the same focal hand, public table state, turn, pass status and seat card counts; the pool is the union of the real opponents' remaining hands. Every combination played by a real opponent belongs to this union and fits its seat's count. The same move is therefore available to the pooled opponent, and removing its cards preserves the mapping. Real passes and forced leads have identical transitions. Induction gives a corresponding pooled trajectory for every real trajectory.

A winning strategy $\phi$ in the pooled game succeeds against all pooled continuations, hence against all real ones on every $d$. The pooled state can be reconstructed from $I$ and the subsequent public actions, so $\phi$ uses no hidden split. Every action in $A^{\mathrm{pool}}(I)$ begins one such strategy. On a fully solved real deal it must therefore belong to $A^{+}(s)$. The inclusion also gives the retention ordering for any distribution over legal actions. □

### A.2 CERTIFIED RESPONSE REDUCTION

We use the complete coalition filtration $\mathcal{F}_t$ and the ordinary-play law of Section 4. The game has bounded payoffs and terminates by $H$; early terminal histories are padded to $H$. At nonterminal eligible histories, the cap $B_t$ is finite and the certificate supplies an executable coalition refutation whose total focal payoff is at most $B_t$. At other times $t < H$ put $B_t = +\infty$. Set $B_H = U$ for terminal boundary notation. An admissible stopping time takes values at eligible histories or at $H$, which means no switch.

The coalition's pooled observations determine the deal, and its history contains the focal move before any post-action switch. Thus eligibility and the selected refutation use information available to the coalition. Couple ordinary play and the switching coalition on the same deal and action randomness until an admissible $\tau$. On $\{\tau = H\}$ their terminal payoffs agree. On $\{\tau < H\}$ the switched payoff $U^\tau$ is at most $B_\tau$, independently of the focal player's subsequent legal choices. With the piecewise completed payoff $C_\tau$ from Section 4,

$$U - U^\tau \geq U - C_\tau, \qquad L_{\mathrm{col}}(\pi) \geq F(\tau) = \mathbb{E}_\pi[U - C_\tau]. \tag{5}$$

The inequality is an equality for this switching coalition when every invoked cap is its exact terminal payoff. Otherwise $F(\tau)$ is the certified lower bound on that coalition's gain.

The original terminal payoff $U$ has the same law for every stopping rule evaluated on the ordinary trajectory. Consequently,

$$F^* = \sup_\tau F(\tau) = \mathbb{E}U - \inf_\tau \mathbb{E}C_\tau.$$

At an eligible history, stopping has completed cost $B_t$, while continuing under the original opponents has conditional cost $\mathbb{E}_\pi[v^*_{t+1} \mid \mathcal{F}_t]$. At an ineligible history only the latter is available. Terminal cost is $U$. Finite backward induction therefore gives exactly (1); choosing its smaller branch constructs an admissible optimal rule. This proves the reduction and $F^* = \mathbb{E}[U - v^*_0]$. It optimizes the bound supplied by the specified certificates.

Intermediate ineligible histories can also be skipped. Condition on the next eligible history or termination, and apply the tower property to the ordinary transitions up to that event. This gives the same recursion on the certificate subsequence, with terminal target $U$. The exact formulation retains a sufficient history and its time index; a coarser feature map still defines an admissible rule, and its evaluated score still satisfies (5).

### A.3 OPTIMALITY INTERVAL AND RESIDUAL ENVELOPE

For $(x)_+ := \max\{x, 0\}$, define the one-sided and absolute envelopes

$$R_\pm(q) := \mathbb{E}_\pi\left[\sum_{t<\tau_q}(-\delta_t)_+ + \sum_{\tau_q \leq t < H}(\delta_t)_+\right], \qquad R(q) := \mathbb{E}_\pi \sum_{t<H}|\delta_t|.$$

The proof below sharpens Theorem 4.1 to $\mathcal{D}(q) \leq R_\pm(q) \leq R(q)$.

*Proof of Theorem 4.1.* Condition on any fitting data so that $q_t$ is a fixed bounded $\mathcal{F}_t$-measurable predictor. Set $v_H = U$, $v_t = \min\{B_t, q_t\}$ for $t < H$, and define

$$d_t = q_t - v_t \geq 0, \qquad \delta_t = q_t - \mathbb{E}_\pi[v_{t+1} \mid \mathcal{F}_t].$$

The process $M$ in Theorem 4.1 is an integrable martingale with $M_0 = 0$. Bounded stopping and optional sampling imply $\mathbb{E}M_\tau = 0$. Therefore

$$\mathbb{E}\min_{t\in\mathcal{T}}(B_t - M_t) \leq \inf_\tau \mathbb{E}C_\tau,$$

the cost-minimization form of the classical stopping dual bound (Rogers, 2002; Haugh & Kogan, 2004). Subtracting from $\mathbb{E}C_{\tau_q}$ proves $F^* - F(\tau_q) \leq \mathcal{D}(q)$, since $\mathbb{E}M_{\tau_q} = 0$. The minimum includes $\tau_q$ on every path, hence $Z_q \geq 0$ pathwise.

For a candidate eligible or terminal time $t$, an exact pathwise identity is

$$B_t - M_t = v_0 + \sum_{s<t}(d_s - \delta_s) + (B_t - v_t).$$

Before $\tau_q$, $d_s = 0$; at $\tau_q$, $C_{\tau_q} - v_{\tau_q} = 0$, including the terminal case. Subtracting candidate times gives

$$Z_q = \max_{t \in \mathcal{T}} \left\{ \sum_{s<t} (\delta_s - d_s) - \sum_{s<\tau_q} \delta_s - (B_t - v_t) \right\}.$$

For $t < \tau_q$, its residual difference is $-\sum_{t \le s < \tau_q} \delta_s$; for $t > \tau_q$, it is $\sum_{\tau_q \le s < t} \delta_s$. The remaining subtracted terms are nonnegative. Consequently, pathwise,

$$0 \le Z_q \le \sum_{s<\tau_q} (-\delta_s)_+ + \sum_{s \ge \tau_q} (\delta_s)_+ \le \sum_{s<H} |\delta_s|.$$

Taking expectations proves $\mathcal{D}(q) \le R_{\pm}(q) \le R(q)$ and the final inequality of (2). At $q = q^*$ every residual is zero, so $Z_{q^*} = 0$ on each path. This paired form evaluates the classical dual gap using common ordinary trajectories.

The residual envelope also follows directly from shared-prefix comparison. For any admissible $\tau$, stopped telescoping and conditional expectation give

$$\mathbb{E}C_\tau - \mathbb{E}v_0 = \mathbb{E}\left[ \sum_{t<\tau} (\mathbb{E}[v_{t+1} \mid \mathcal{F}_t] - v_t) + \mathbf{1}\{\tau < H\}(B_\tau - v_\tau) \right] \tag{6}$$

$$= \mathbb{E}\left[ \sum_{t<\tau} (d_t - \delta_t) + \mathbf{1}\{\tau < H\}(B_\tau - v_\tau) \right]. \tag{7}$$

Indeed, $\mathbf{1}\{t < \tau\}$ is $\mathcal{F}_t$-measurable, so each expected increment can be replaced by its conditional expectation. At $\tau = H$ the terminal difference vanishes because $B_H = v_H = U$. At a preterminal stop the boundary difference is nonnegative because $v_\tau \le B_\tau$. The finite horizon and bounded quantities justify the sums and expectations.

For the greedy rule $\tau_q$, continuation implies $v_t = q_t$, hence $d_t = 0$, and stopping implies $v_{\tau_q} = B_{\tau_q}$. Its two nonnegative gap terms therefore vanish:

$$\mathbb{E}C_{\tau_q} = \mathbb{E}v_0 - \mathbb{E}\sum_{t<\tau_q} \delta_t.$$

For the optimal rule $\tau^*$, retain the nonnegative terms in (7) and subtract on the same ordinary trajectory. This yields

$$\begin{aligned} 0 \le F^* - F(\tau_q) &= \mathbb{E}C_{\tau_q} - \mathbb{E}C_{\tau^*} \\ &\le \mathbb{E}_\pi \sum_{t=0}^{H-1} \left( \mathbf{1}\{t < \tau^*\} - \mathbf{1}\{t < \tau_q\} \right) \delta_t \\ &\le \mathbb{E}_\pi \sum_{t=0}^{H-1} |\delta_t|. \end{aligned}$$

The prefix-indicator difference is zero before both stopping times and after both stopping times. Between them its magnitude is one. The coefficient therefore comes from a single residual sum on the original trajectory.

For sharpness, consider one eligible history with $B = -1$ and deterministic ordinary terminal payoff $U = 3$. Set $q = -1 - \varepsilon$ for $\varepsilon > 0$, and use exact continuation values thereafter. The fitted rule continues while the optimal rule stops. Its regret is four and its residual norm is $4 + \varepsilon$, so their ratio tends to one. Thus no uniformly smaller coefficient is possible.

This example occurs in Big Two v2. The focal seat holds $3\diamondsuit$ and owns a $5\diamondsuit$ table; the next seat holds $6\diamondsuit$ and the remaining seats hold $4\diamondsuit$ and $4\clubsuit$. The next opponent can finish immediately. Under ordinary play it instead passes, as do the next two seats, leaving the focal seat to lead its last card and collect three. The certified denial plays $6\diamondsuit$ and gives the focal seat exactly minus one. □

The residual is a conditional-mean quantity. A sampled temporal-difference error also contains transition noise; its squared average is a regression diagnostic, not the norm in (2).

### A.4 EXACT LAST-CARD BOUNDARY

**Theorem A.2** (Exact value of last-card denials)**.** *In Big Two v2, let every eligible trigger occur before the first finish, after a focal action that leaves one card. Write $T_g = \infty$ when rule $g$ never triggers, and let $W$ be first place on the original trajectory. The pathwise equality* (3) *holds, and*

$$\mathbb{P}_\pi(W) - \mathbb{P}_{\pi,\sigma_g}(W) = \mathbb{P}_\pi(T_g < \infty, W). \tag{8}$$

*Its exact gain $F_g = \mathbb{E}[U - U^g]$ satisfies*

$$L_{\mathrm{col}}(\pi) \geq F_g = \mathbb{P}(T_g < \infty)\mathbb{P}(W \mid T_g < \infty)\big(1 + \mathbb{E}[U \mid T_g < \infty, W]\big). \tag{9}$$

*The product is zero when its conditioning event has probability zero. Let $\mathcal{E}_g$ denote the set of histories at which rule $g$ may trigger. For nested eligibility sets $\mathcal{E}_g \subseteq \mathcal{E}_{g'}$, the per-game gain and its expectation are nondecreasing. If at most $m$ opponent cards remain at every eligible trigger, the score belongs to $\{0, 4, \ldots, m+1\}$.*

*Proof.* At a trigger the focal hand contains one card and no player has finished. Cards are only removed. Playing that card would make the focal player the first finisher, which the certified denial prevents against every focal continuation. The focal player therefore retains exactly one card when another player finishes. Big Two v2 settles then, with each non-winner paying its retained card count, so $U^g = -1$ on every switched trajectory.

On the original trajectory, the focal player either also retains that card until another player finishes, paying $U = -1$, or wins and receives the sum of the other three retained hand sizes. Those hands are nonempty at the first finish. Their total cannot increase after the trigger, so $3 \leq U \leq m$ on the winning event. Non-triggered trajectories coincide. This proves (3), its nonnegativity, and the stated support. The first-place indicator difference is exactly $\mathbf{1}\{T_g < \infty, W\}$, proving (8). Taking expectations and conditioning on the trigger and original win proves (9).

For nested eligibility sets on an original trajectory, occurrence of the smaller-set stopping event implies occurrence of the larger-set event. When both occur, the two gains use the same original terminal $U$ and switched payoff $-1$, even if their stopping times differ. When only the larger-set event occurs, its gain is nonnegative. The ordering therefore holds pathwise. Any selective or delayed rule in this last-card class likewise triggers on a subset of the trajectories reached by the first-available rule. The same argument proves optimality of the first-available rule within the class; randomized selection follows by conditioning on its randomness.

Finally, from a last-card history every subsequent ordinary payoff is at least $-1$, and each valid refutation cap at a reachable certified history is also at least $-1$. Backward induction gives $q_t^* \geq -1 = B_t$ at the last-card boundary. Thus immediate switching is optimal in the completed-payoff recursion as well. □

### A.5 CONTINUATION ERROR AND BELLMAN RESIDUAL

To emphasize the dependence on the supplied predictor, write its residual as $\delta_t(q) := q_t - \mathbb{E}_\pi[v_{t+1} \mid \mathcal{F}_t]$. In undiscounted cost notation, Haugh & Kogan (2004, Theorem 3) control the greedy stopping rule by

$$F^* - F(\tau_q) \leq \mathbb{E} \sum_{t=0}^{H-1} |q_t - q_t^*|.$$

Their right-hand side uses the optimal continuation function. The residual envelope in Theorem 4.1 instead uses the supplied predictor's conditional self-consistency. Define $e_t := q_t - q_t^*$ for $t < H$ and $e_H := 0$. The Bellman recursion and the 1-Lipschitz property of $\min\{B, \cdot\}$ give

$$\mathbb{E}|e_t| \leq \mathbb{E}|\delta_t(q)| + \mathbb{E}|e_{t+1}|, \qquad e_H = 0.$$

Substitution into the continuation-error bound would therefore give

$$F^* - F(\tau_q) \leq \sum_{s=0}^{H-1} (s+1)\mathbb{E}|\delta_s(q)|.$$

The shared-prefix proof removes this repeated counting and gives $R(q)$ with coefficient one. The same authors' martingale-dual construction supplies the finer certificate $\mathcal{D}(q) \le R_{\pm}(q) \le R(q)$ established above. Thus the continuation-error comparison and the dual comparison concern distinct guarantees; the measured optimality certificate in this paper uses the latter.

For a concrete separation, take a deterministic chain with $H$ nonterminal decision times, all eligible. Set the ordinary terminal cost to $U = -3$, the caps to $B_t = -1$ for $t < H-1$ and $B_{H-1} = -2$, and the predictor to $q_t = -2$ throughout. Never switching is optimal, so $q^*_t = -3$. The greedy predictor waits until the last decision and switches there, giving regret one. Its residual is zero until that last decision and one there: the residual bound is exactly one, whereas $\sum_t |q_t - q^*_t| = H$. Every time permits stopping, so suppressing ineligible times does not shorten this example. The two bounds use different errors; this example establishes a horizon-dependent separation for the displayed family.

Bhandari & Russo (2024, Section 7) instead study a discounted stationary stopping process and residuals of policy cost-to-go functions. Their favorable occupancy coefficient comes from the never-stopping process. Our finite-horizon statement concerns arbitrary adapted continuation predictors, with terminal condition $v_H = U$, and controls the regret of their greedy stopping rule under the original full-game law. The stopping recursion and favorable coverage are established ingredients; the certificate boundary and the residual comparison specify the construction used here.

**Localizing a measured residual.** The same proof also gives

$$F^* - F(\tau_q) \le \mathbb{E} \sum_{t=0}^{H-1} \mathbf{1}\{\min(\tau_q, \tau^*) \le t < \max(\tau_q, \tau^*)\} |\delta_t(q)|.$$

We use this quantity as an oracle-dependent diagnostic when $\tau^*$ is available by exact evaluation. It identifies error in the portion of the ordinary path on which the two stopping rules disagree. The deployable predictor's global residual remains the bound in Theorem 4.1. In particular, assigning $q = B = -1$ at a last-card boundary gives the correct stopping action even when the ordinary continuation value is larger; a global residual can include that discrepancy after stopping.

The dual score has a useful boundary invariance. Changing $q_t$ only at eligible histories where both its old and new values are at least $B_t$ leaves $v$, the greedy rule, $M$ and $Z_q$ unchanged. The raw residual can change through its $q_t$ term. Values below the cap also determine continuation, so this invariance concerns the stopping region specifically.

**Independent successor simulation.** Following nested dual evaluation (Andersen & Broadie, 2004), replace the conditional next value by an average $\widehat{m}_s$ of $k$ fresh ordinary successors drawn independently of the primary successor and its future. Let

$$\mathcal{G}_s = \mathcal{F}_s \vee \sigma(\text{auxiliary draws at times } 0, \dots, s-1), \qquad \widehat{M}_t = \sum_{s<t} (v_{s+1} - \widehat{m}_s).$$

The current auxiliary draws are revealed together with the next primary history. Conditional on $\mathcal{G}_s$, both terms in the increment have mean $\mathbb{E}[v_{s+1} \mid \mathcal{F}_s]$, so $\widehat{M}$ is a martingale. The original $\tau_q$ remains adapted. Define

$$Z_{q,k} := C_{\tau_q} - \widehat{M}_{\tau_q} - \min_{t \in \mathcal{T}} (B_t - \widehat{M}_t), \qquad \mathcal{D}_k(q) := \mathbb{E}_\pi Z_{q,k}.$$

The same optional-sampling argument shows that $\mathcal{D}_k(q)$ is an expected regret bound.

Conditional on a complete primary path, the auxiliary errors have mean zero. Concavity of the pathwise minimum gives $\mathcal{D}_k(q) \ge \mathcal{D}(q)$. The pathwise residual comparison uses the sampled residual $q_s - \widehat{m}_s$ and its own one-sided sum. It does not compare $\mathcal{D}_k$ to the exact-conditional $R_{\pm}$. Finite inner sampling can consequently give a positive gap even at $q = q^*$. We separate this inner-sampling effect from outer-path estimation uncertainty; neither a sample mean nor a single path is itself an expected-value confidence bound.

The same martingale can certify another admissible rule $\rho$. Replace the chosen-stop term by $C_\rho - \widehat{M}_\rho$ to obtain

$$Z_{\rho,q,k} = C_\rho - \widehat{M}_\rho - \min_{t \in \mathcal{T}} (B_t - \widehat{M}_t) \ge 0, \qquad F^* - F(\rho) \le \mathbb{E} Z_{\rho,q,k}.$$

Optional sampling and the same pathwise minimum prove this statement; the Bellman-residual envelope in Theorem 4.1 concerns the specific choice $\rho = \tau_q$. For randomized forecast gates, their

internal draws are independent of the ordinary continuation and the auxiliary dual draws, conditional on current information. Including past and current gate draws in the filtration preserves the martingale. A learned continuation can therefore supply the dual certificate for a separately simulated switching rule.

### A.6 Absorbing a certified boundary

An exact tactical boundary can remove uncertainty from dual evaluation as well as from the deployed switch. The required property is domination of all later cap costs, including the terminal ordinary payoff.

**Proposition A.3** (Boundary absorption)**.** *Let $T \le H$ be a stopping time that is eligible whenever $T < H$. Suppose, pathwise, $B_s \ge B_T$ for every eligible $s \ge T$, including $s = H$. Let $q_T \ge B_T$ whenever $T < H$. Replace the ordinary stopping process after $T$ by a terminal boundary of value $B_T$, leaving the original ordinary payoff $U$ as the reference in $F$. Let $\widetilde{Z}_q$ denote the paired dual score of this absorbed process. Then the optimal cap-valued gain and the gain of $\tau_q$ are unchanged. Coupling the original and absorbed duals on their common prefix gives*

$$0 \le \widetilde{Z}_q \le Z_q \quad \text{on every path}, \qquad F^* - F(\tau_q) \le \mathbb{E}\widetilde{Z}_q.$$

*The same ordering and expected regret bound hold for independent-successor duals when their auxiliary draws are coupled before $T$.*

*Proof.* For any admissible $\tau$, the rule $\tau \wedge T$ is admissible and $C_{\tau\wedge T} \le C_\tau$ pathwise. Restricting to $\tau \le T$ therefore preserves the infimum of expected cap cost. The rule $\tau_q$ already belongs to this class because it stops at $T$ if it has not stopped earlier. The reference $\mathbb{E}U$ is unchanged, establishing both gain identities.

The original fitted value at $T$ is $v_T = B_T$, the terminal value in the absorbed process. Consequently the next-value conditional expectations, and hence the martingale increments, agree before $T$. The chosen-stop term $C_{\tau_q} - M_{\tau_q}$ is identical. The absorbed pathwise minimum ranges only over eligible times at most $T$, a subset of the original candidates, so it is at least the original minimum. Subtraction gives the pathwise ordering. Stopping duality in the absorbed problem gives the regret bound. With independent successors, use the same draws for the two constructions before $T$; their sampled next values agree, and the enlarged-filtration martingale argument applies to each. The same minimum comparison then proves the sampled ordering. □

For Big Two v2, take $T$ to be the first eligible denial with one retained focal card, or $H$ if none occurs. Later eligible caps equal $-1$, and the ordinary payoff is at least $-1$ once that card count is reached. Thus the proposition applies to every last-card-anchored gate. Equivalently, extend the absorbed martingale to the original game as $\widetilde{M}_t = M_{t\wedge T}$. Every later adjusted cap is at least $-1 - M_T$, already an available boundary, so the original game's dual minimum equals the absorbed minimum. The ordinary game still completes to measure $U - C_{\tau_q}$; only its optimality-certificate simulation is absorbed. No unobserved suffix outcome is used to decide where to absorb.

### A.7 Deterministic continuation oracle

Suppose ordinary policies and game transitions are deterministic conditional on the coalition's current history, which contains the state needed to execute those policies. Let all eligible boundaries use the same finite cap $b$, and let $\mathcal{A}$ denote existence of an eligible boundary on the ordinary trajectory. At the first such boundary, simulating the ordinary continuation computes its terminal payoff $U$ using only that history. Thus the rule that switches there when $U > b$ and otherwise never switches is admissible.

On every ordinary trajectory, a switch at any eligible time has score $U - b$ and no switch has score zero. The largest possible cap-valued score is therefore $\mathbf{1}\{\mathcal{A}\}(U - b)_+$. The stated rule attains it, proving (4). This is the optimum for the given eligibility predicates and constant payoff cap. No property of a more general coordinated response is used.

For deterministic DMC, the coalition knows all hands and the public move history consumed by the frozen policy, and Big Two v2 has no within-game chance event. Ordinary rollout is consequently

an available computation at the switch. The implementation forecasts from a copy of the complete state after the observed denial action, including the public history, and compares the forecast with $-1$. Its forecasting function receives no stored terminal reward.

In the measured regime, the optimum score is in $[0, 14]$. The paired regret relative to any other rule in the same constant-cap class is also in $[0, 14]$: when $U+1 > 0$, regret is either zero or $U+1 \le 14$; when $U + 1 \le 0$, it is either zero or $-(U + 1) \le 7$. This sharper paired support follows from the common trajectory and cap.

At the ordered eligible times of deterministic play, the conditional expectation in the Bellman residual equals the next observed value. Hence it can be evaluated exactly as $q_t - \min\{-1, q_{t+1}\}$ at a nonfinal eligible time, and $q_t - U$ at the last one. For this compressed time index, $q_t^* = \min\{-1, U\}$ if another eligible time remains and $q_t^* = U$ otherwise. An initial transition before the first eligible time can be extended with its exact continuation and zero residual. These identities provide full-game residual and optimality-gap measurements for the frozen predictors. For the DMC window diagnostic we use the optimal forecast rule above: first denial when $U > -1$, and no switch when $U \le -1$. This fixes the otherwise arbitrary optimal tie choice at $U = -1$.

## B FINITE-SAMPLE CONFIDENCE CERTIFICATE

**Corollary B.1** (Simultaneous confidence certificate)**.** *Fix $K$ stopping rules independently of $n$ identically distributed, independent evaluation games, and finish all $n$ games. For each rule index $g \in \{1, \dots, K\}$, let $X_g = U - C_{\tau_g}$ be its score, with known finite support $z_{g,0} < \cdots < z_{g,m_g}$. For a target error probability $\alpha \in (0, 1)$, set $M = \sum_{g=1}^{K} m_g > 0$. Let $\ell_{n,\delta}(k)$ be the one-sided Clopper–Pearson binomial lower limit at error level $\delta \in (0, 1)$ (Clopper & Pearson, 1934). For $k_{g,j} = \sum_{i=1}^{n} \mathbf{1}\{X_{g,i} \ge z_{g,j}\}$, define*

$$\underline{F}_g = z_{g,0} + \sum_{j=1}^{m_g} (z_{g,j} - z_{g,j-1}) \ell_{n,\alpha/M}(k_{g,j}). \tag{10}$$

*With probability at least $1 - \alpha$, simultaneously $\mathbb{E}X_g \ge \underline{F}_g$ for every rule and $L_{\mathrm{col}}(\pi) \ge \max_g \underline{F}_g$. Constant scores have known means and require no tail tests.*

*Proof.* For $k > 0$ define $\ell_{n,\delta}(k) = \mathrm{Beta}^{-1}(\delta; k, n - k + 1)$ and set $\ell_{n,\delta}(0) = 0$. Equivalently, for $k > 0$ the lower limit is the value $p_0 \in [0, 1]$ that solves $\mathbb{P}_{p_0}\{\mathrm{Bin}(n, p_0) \ge k\} = \delta$. Monotonicity of the binomial tail gives

$$\mathbb{P}\{\ell_{n,\delta}(K_{\mathrm{bin}}) > p\} \le \delta \quad \text{for } K_{\mathrm{bin}} \sim \mathrm{Bin}(n, p).$$

For each fixed rule and threshold, the complete-game indicators $\mathbf{1}\{X_{g,i} \ge z_{g,j}\}$ are identically distributed, independent Bernoulli observations. Apply this bound at level $\alpha/M$ and take a union bound over the $M$ rule-threshold pairs. All tail lower limits then hold simultaneously. Dependence between different rules or thresholds on the same game does not affect the union bound.

For every possible score,

$$X_g = z_{g,0} + \sum_{j=1}^{m_g} (z_{g,j} - z_{g,j-1}) \mathbf{1}\{X_g \ge z_{g,j}\}. \tag{11}$$

Taking expectations and substituting the simultaneous binomial limits proves the mean bounds. Each rule's mean is at most $L_{\mathrm{col}}$ by (5), so their maximum also supplies a valid lower bound. For rules fitted on separate data, first condition on that data and then average the probability guarantee over it. □

Signed general scores and exact nonnegative last-card scores can share the same evaluation family. Supports are fixed from the game rules and eligibility conditions, rather than from observed extrema. The same tail construction applies to a paired difference $X_g - X_{g'}$ with known finite support; including its thresholds in $M$ certifies the mean improvement simultaneously with the individual gains. Every complete-game score, including negative values and games with no trigger, remains in the sample.

# C ENVIRONMENTS AND MEASUREMENT PROTOCOLS

**Game rules and public policies.** Both Big Two variants use four seats and a fully dealt 52-card deck. The v1 engine follows the Charlesworth PPO release; v2 follows big2-rl's ranking and legal-combination rules. In v2 the first finish ends the game, each losing seat pays its number of remaining cards, and the winner receives their sum. The reference implementation's multiplier thresholds exceed the maximum 13-card hand, so no multiplier applies in these experiments. TLMN also uses four seats and all 52 cards, with singles, groups, runs and chops. Its engine settles in ticks, six ticks per stake; full-game evaluation divides focal rewards by six. The payoff theorem's measured exact last-card instance is Big Two v2.

The released DMC policy acts by deterministic argmax; the PPO release and our learned policies use their sampling behavior in the own-play and refutation audits. The DMC bridge matches its reference action on $2{,}553/2{,}553$ tested decisions. The PPO bridge uses the release's observation and action encoding. Training-set retention and full-game continuation modes are stated separately. New retained-action measurements score greedy and sampled behavior explicitly, so a probability-mass claim never relies on a greedy action alone.

**Solver interface.** Paranoid search quantifies over every focal action and every joint opponent continuation on the actual deal. Pooled search replaces the hidden split by the unseen union while retaining turn, counts, table owner and pass state. Search depth is unbounded; a finite node cap returns unknown without assigning a win or loss. Big Two v2 audits use exhaustive Python search across 24 workers. The TLMN training-pool builder uses a CUDA implementation, checked against the Python solver on lead and mid-trick states. Both the action before a throw and the denial after it must be proved. This is different from a depth-limited tactical label.

Before the new payoff audits, a fresh GPU/reference gate verifies 512 complete games and 39,531 decisions with exact legality and reward agreement. Independent reference-search comparisons cover 40 lead and 40 response states; pooled soundness checks enumerate hidden splits for 25 lead and 25 response states. A fixed-cohort test with 31 games and unequal lane quotas checks that every assigned game completes and no in-flight game is excluded. These tests cover state transitions, solver verdicts and cohort inclusion separately.

**Own-play errors and refutation.** The recorded own-play decisions are sampled from each policy's rollout inside its solver regime. Conditional throw rates divide proved throws by proved-win decisions in that sample; unknowns remain a separate category. For each recorded throw, the same move is forced at the same saved state in the mirror and refutation arms. Subsequent focal choices follow the audited behavior mode, and common random numbers couple comparable choices. All five rows have zero refutation first-place rate and zero fallbacks. These paired state-level replays are separate from the full-game training matchups.

| Population | Train | Held-out | Selection and overlap check |
|---|---|---|---|
| Big Two natural | 5,266 | 5,289 | Self-play; zero information-key overlap |
| TLMN natural pilot | 2,404 | 2,422 | Self-play; used here as a separate probe |
| TLMN challenge | 373 | 195 | Templates; zero key and focal-hand overlap |

Table 6: Frozen certified populations. Big Two allows repeated focal hands with different public information (655 hand overlaps across splits). TLMN challenge hands are unique within and across the two sets.

**Challenge-set construction.** Big Two training and held-out examples come from a natural self-play harvest with pooled winning actions. Each position has both true-deal and pooled action sets. TLMN's challenge generator first stratifies templates by whether a frozen reference's greedy action retains the template's bounded-depth win, then applies unbounded-depth paranoid and pooled certification. The 195-position held-out challenge set belongs to that template-failure stratum. Final certification supplies $A^{\mathrm{pool}}$. The same reference attains greedy $\rho_{\mathrm{deal}} = 189/195$ and $\rho_{\mathrm{pool}} = 162/195$ on these final labels. Thus the challenge selection is defined by template difficulty, and the reported final metrics use the stronger certified action sets. The separate 972-candidate solver-port decomposition and earlier 64-position template probe are different populations.

**Reset resampling.** At each injected reset, shuffle the unseen cards into hands with the stored per-seat counts; preserve the focal hand, unseen union and public table. The certificate is valid for every such hidden split. A reference gate enumerates 90 splits and tests 1.8 million draws (maximum relative frequency deviation $1.8\%$); full-pool checks preserve all visible-state invariants. The sampled split defines the curriculum distribution. The soundness proof does not assume that this distribution equals the on-policy posterior.

**Training and additional evaluation.** The paired training comparisons use fixed final-checkpoint selection. The TLMN comparison is specified before either arm launches, and its training/held-out split is fixed before the certified arm consumes it. Control training does not use that pool. The additional two-label/two-mode and complete-cohort evaluations apply to these frozen final policies.

**Retention and uncertainty.** The action audit uses 128 draws per state and common random numbers across paired policies. It reproduces the original greedy metrics within $10^{-6}$ and checks $A^{\text{pool}} \subseteq A^{+}$ and the resulting per-state retention ordering. Let $\Delta_i$ be the certified-minus-control difference for training-run pair $i$, and let $\bar{\Delta}$ and $s_\Delta$ be the sample mean and standard deviation of the five differences. We write $t_{d,p}$ for the $p$ quantile of Student's $t$ distribution with $d$ degrees of freedom. The paired intervals in Table 5 are $\bar{\Delta} \pm t_{4,0.975} s_\Delta / \sqrt{5}$. These intervals describe variation across training runs on the fixed benchmark. The TLMN two-way bootstrap independently resamples the five paired runs and 195 positions, preserving control/certified pairing and averaging over their Cartesian product. Its 20,000 replicates include variation across observed challenge positions and training runs.

**Complete-cohort training evaluation.** A matchup places one candidate against three copies of its paired-run control, all at native sampling temperature. Each of five final-checkpoint pairs receives the same 200,000 assigned games, balanced exactly across the four candidate seats. A wave assigns one game per lane and completes every lane before advancing; no lane redeals within a wave. Deal and per-ply action streams are fixed, so completion time cannot change another game's random draws. Every Big Two control comparison uses the same initial cohort. TLMN includes all 5,961 games per pair that settle at the initial deal, with their initial rewards and zero plies. No assigned game is excluded. Training-run intervals condition on this shared cohort. Game SE instead averages the five policy-pair scores within each common game and combines independent seat strata, conditional on the frozen policies. Individual deals are not replayed in all seats.

## D WORKED EXAMPLES FROM RECORDED GAMES

This appendix shows one recorded position from each game and what the audit computes there. Seats are numbered in turn order. In Big Two and TLMN, P0 is the focal player and the other three seats form the coalition; in Guandan, one partnership is audited and the other responds. Every displayed move is recorded play, read from stored trajectories or regenerated by a replay that reproduces the stored audit results exactly. Suits follow each game's own order.

**Big Two v2: a denial that ordinary play forgives, and the switching decision**

**Position.** Deal 2 of the independent 60,000-game cohort in Table 2. It is the first deal with an eligible denial; all four seats play the released DMC policy. P2 has won the lead with the pair [7♢] [7♡], P3 and the focal P0 have passed, and P1 is to act.

| | |
|---|---|
| P0 (focal) | [8♠] [10♢] |
| P1 | [3♡] [7♠] [10♡] [Q♣] |
| P2 | [6♣] [9♣] [10♠] [K♢] [2♣] |
| P3 | [5♠] [6♡] [K♠] |

**Certificate.** Exhaustive search proves that, after P0's pass, the other three seats can stop P0 from finishing first against every later focal choice. Before the first finish, this caps the focal payoff at $B = -1$.

**Ordinary continuation.** P1 pass, P2 [6♣], P3 [6♡], P0 [8♠], P1 [10♡], P2 [10♠], P3 [K♠], two passes, P2 [2♣], three passes, P2 [9♣], P3 pass, P0 [10♢]. P0 goes out first and collects the five cards left in the other hands: $U = +5$.

**Executed refutation.** The first 13 moves are identical. P2 then leads [K♢], which no remaining card beats, and finishes with [9♣] after three passes. P0 still holds [10♢] and pays $-1$.

**Switching.** The coalition sees all four hands and plays against a deterministic policy, so it can forecast $U = +5$ from its current history. Since $U > B$, the oracle of Equation (4) switches at this denial with

certified gain $U - B = 6$, and the executed refutation attains the cap exactly. The first five deals show how every game, including those without a switch, enters the averages of Table 2:

| Deal | Eligible denial | Forecast $U$ | Oracle action | Oracle | Anchored | Monte Carlo | Last-card |
|---|---|---|---|---|---|---|---|
| 0 | no | — | — | 0 | 0 | 0 | 0 |
| 1 | no | — | — | 0 | 0 | 0 | 0 |
| 2 | yes | $+5$ | switch | 6 | 6 | 6 | 6 |
| 3 | yes | $-4$ | continue | 0 | 0 | 0 | 0 |
| 4 | yes | $-1$ | continue | 0 | 0 | 0 | 0 |

**Big Two v1: the same audit under the original rules and a sampled PPO release**

**Differences from v2.** Big Two v1 follows the released PPO implementation. It adds four-card plays (two pair and four of a kind) and allows straights only from 3-4-5-6-7 up to J-Q-K-A-2. Big Two v2 has no four-card plays, scores four of a kind as a five-card hand, and also counts A-2-3-4-5 and 2-3-4-5-6 as straights. Both variants pay the winner the cards left in the other hands. The audited v1 policy samples its actions, whereas the v2 DMC policy acts by argmax.

**Position.** One of the 160 recorded throws in Table 1. Mirror play forgives 43 of them; we show the forgiven throw with the fewest cards in play. P3 has just played 6♡, and the focal P0 must answer. The shaded card is the certified winning action, and the dashed card is the policy's choice.

P0 (focal) 8♢ 10♢
P1 7♠
P2 3♣ 3♠
P3 3♡ 9♠ J♢ 2♢

**True-deal audit.** Search proves that 8♢ is the only winning answer and that 10♢ loses; the policy played 10♢. After 8♢, P3 can take the trick only with J♢ or 2♢, because 9♠ would let 10♢ out at once. It would then still need two unbeaten leads before its last card, with only one such card left. After 10♢, both 9♠ and J♢ beat the remaining 8♢.

**Mirror continuation.** P3 beats 10♢ with J♢, leads 9♠, and then leads 3♡ while still holding 2♢. P0 answers with 8♢ and collects $U = +4$. **Executed refutation.** P3 beats 10♢ with 2♢, leads J♢ and 9♠, each followed by three passes, and finishes with 3♡. P0 pays $-1$.

**Comparison.** As in the v2 example, ordinary play forgives because an opponent gives up the lead one card too early. The audits differ in solver regime: the node-capped v1 search leaves 6.0% of candidate decisions unresolved, while the v2 search is exhaustive.

**TLMN: a throw, its information-state certificate, and the two continuations**

**Position.** One of the 232 recorded throws of the frozen TLMN reference in Table 1, sampled at its training temperature. Mirror play forgives 55 of these throws; we show the forgiven throw with the fewest cards in play whose information state has a pooled certificate. The focal P0 leads to an empty table. The shaded card is the certified winning action, and the dashed card is the policy's choice.

P0 (focal) 7♢ 2♣
P1 10♡ J♣ J♢ Q♠ K♡
P2 10♠
P3 Q♢ K♢ A♢

**True-deal audit.** Exact search proves that leading 2♣ wins and leading 7♢ loses, so $A^+(s)$ and $A^-(s)$ each contain one action. The policy led 7♢: a throw.

**Information-state certificate.** P0 sees its own hand, every played card and the counts $5, 1, 3$. It therefore knows the union of the nine hidden cards, but not their split. The pooled solver still certifies that 2♣ is the only action in $A^{\text{pool}}(I)$. It is the last two in play, and a single two can be chopped only by a quad or by three or four consecutive pairs. The pool holds pairs of tens, jacks, queens and kings, but no hidden seat has the six cards such a run needs, and no rank has four copies. Hence 2♣ wins the trick for every split, and 7♢ then goes out. The throw is information-chargeable.

**Mirror continuation.** P1 Q♠, P2 pass, P3 A♢, P0 2♣: P0 finishes first. **Executed refutation.** P1 pass, P2 10♠: the one-card seat finishes first, and P0 is denied.

**Guandan: a partnership certificate that holds for every hidden allocation**

**Position.** From the 20,000-round natural-state evaluation of Appendix I. The audited partnership wins 4 of its 644 certified rounds under ordinary play; this is the first of those four. The level rank is six, so sixes rank above aces. Partnership V (P0, P2) is audited and partnership C (P1, P3) responds. P3 has played the bomb 2♣ 2♡ 2♡ 2♠, P0 has passed, and P1 is to act.

P0 (V) 7♠ P1 (C) 10♣ Q♣
P2 (V) 10♠ 6♣ P3 (C) 5♣ 5♡ 6♠ J♣

**What C knows.** C sees its two hands and the union 7♠ 10♠ 6♣ of V's cards, with counts 1 and 2. It does not know which V seat holds which card.
**Certified witness.** After P1 and P2 pass, P3 leads the pair 5♣ 5♡; the union contains no pair. It then leads the level card 6♠, which nothing in the union beats. It finishes with J♣, the one card that 6♣ could have beaten had it been led earlier. V passes throughout, so C finishes first under either allocation, and V's payoff is at most $-1$.
**Outcome.** The executed round ends at $-2$ for V, within the cap. Under ordinary play, V had won this round with $+1$.

## E FULL-GAME PAYOFF AUDIT DETAILS

The initial payoff audit covers 360,000 games: 60,000 for DMC and 30,000 for each of ten trained checkpoints. It assigns lane and episode quotas before rollout, generates exactly the assigned number of games for each policy, and completes them all. Deals and action randomness are fixed before outcomes. All policies share the first 30,000 initial deals, and focal seat 0 is fixed. For sampled policies, randomness is indexed by deal identifier and ply; a different game length cannot shift subsequent deals or another game's action stream. The estimator averages one score per full game, including zeros and signed negative scores. Common deals couple policy comparisons, while the per-policy confidence theorem uses independence between games.

The five eligibility predicates are fixed before evaluation. First true-deal and first information-chargeable throws each search for their own earliest qualifying event. Their score support is contained in the integer range $[-7, 14]$: pre-move focal hands have at most eight cards and at most thirteen opponent cards remain. The three last-card rules have support contained in $[0, 14]$. We use all integer thresholds, including impossible values, for a conservative fixed family of $21 + 21 + 14 + 14 + 14 = 84$ binomial tails. Set the overall error probability to $\alpha = 0.05$. Allocating $\alpha/84$ per tail gives simultaneous bounds for the five rules at one frozen checkpoint. Using $\alpha/(11 \cdot 84)$ additionally makes the family simultaneous across the eleven audited checkpoints.

Across all eleven checkpoints and five rules, the simultaneous 95% lower bound for DMC's exact last-card denial value is 0.29665. The corresponding lower bounds for the ten trained policies range from 0.10397 to 0.11446. DMC uses 60,000 complete games; every trained policy uses the same 30,000 deals. All bounds use one joint family of 924 binomial tails. The broad denial values average

| Rule | Control | Certified | Paired change [95% CI] |
|---|---|---|---|
| First true-deal throw | 0.03086 | 0.02990 | $-0.00096$ $[-0.00471, +0.00279]$ |
| First information-chargeable throw | 0.00217 | 0.00177 | $-0.00040$ $[-0.00171, +0.00091]$ |
| Last-card true-deal throw | 0.00218 | 0.00149 | $-0.00069$ $[-0.00125, -0.00014]$ |
| Last-card information-chargeable throw | 0.00075 | 0.00063 | $-0.00012$ $[-0.00080, +0.00056]$ |
| Last-card denial | 0.13640 | 0.13548 | $-0.00092$ $[-0.00662, +0.00478]$ |

Table 7: Five-run means and paired training-run differences at the common 30,000-deal audit. Intervals are unadjusted descriptive Student-$t$ intervals for each contrast, separate from the simultaneous population lower bounds.

0.13640 for controls and 0.13548 for certified policies. Their paired difference is $-0.00092$ with interval $[-0.00662, 0.00478]$. The last-card throw component decreases by 0.00069 per game, while broad denial exposure remains similar. These measure specific switching strategies, so a smaller score alone is not an upper bound on overall coalition vulnerability.

The DMC witness check takes the first 256 eligible games without inspecting original terminal payoffs. A random legal focal continuation tests an executable denial in an exact transition model; the

denial invariant is checked at every node. All switched payoffs are $-1$. The separate GPU/reference check verifies the settlement model used by the ordinary rollout.

# F FROZEN LEARNED STOPPING PROTOCOL

**Development and independent evaluation.** We use DMC's initial $60{,}000$ complete games as development data, with $48{,}000$ games for fitting and $12{,}000$ disjoint games for validation. The ordered denial opportunities within each game form the continuation transitions. Their $125$ features comprise current hands and card counts, the table, its owner, pass count, ply, chosen action and current certificate verdicts. All are available from the coalition's joint information when the focal action is observed. Terminal rewards and later states supply training targets and evaluation scores only. This representation omits public move-history inputs consumed by DMC; regret relative to its full-history optimum includes both representation and fitting error. The stochastic comparison below instead uses a representation sufficient for its frozen policy and transitions.

After selecting and freezing both predictors and their thresholds, we generate a separate cohort of $60{,}000$ complete DMC games with independent deal randomization. Its size is fixed before play; it shares no initial deal with the development population. All eight rules are evaluated on this same cohort. The candidate regime has at most eight focal cards, thirteen opponent cards in total and no prior finish. The four throw-based rules retain their original lower limit of two focal cards; denial-based rules also admit one-card hands. Each rule uses its own first qualifying event.

**Continuation fitting and outcome prediction.** The first anchored continuation fit evaluates the last-card rule: its target is $-1$ when a later last-card denial occurs, and the ordinary terminal payoff $U$ otherwise. Seven subsequent fitting stages use $\min\{-1, q_{\mathrm{old}}(I'_C)\}$ at the next denial, or $U$ if no denial remains. The known last-card boundary is always assigned value $-1$. Each stage fits a squared-error histogram gradient-boosted regressor with $64$ trees, learning rate $0.08$, at most $31$ leaves per tree, at least $80$ observations per leaf, $\ell_2$ regularization $1$, and at most $63$ bins per feature. Early stopping is disabled. The Monte Carlo predictor uses the same features and parameters, $512$ trees, and ordinary terminal payoff targets; each fitting procedure therefore uses $512$ boosting steps in total. Predictions are clipped to $[-8, 13]$.

Both learned rules stop immediately at a last-card denial. At other denials they stop when $q(I_C) \geq -1+\eta$. Validation chooses the fitting stage and margin from $\eta \in \{0, 0.1, 0.25, 0.5, 1\}$ by the largest mean signed game score, with ties resolved by the earlier stage and then the smaller margin. It selects the first anchored stage with $\eta = 0$ and the Monte Carlo predictor with $\eta = 0.5$. The selected stage is a $64$-tree continuation predictor; the outcome predictor has $512$ trees. Fitting and selection take $24.9$ seconds. The independent rollout takes $14.9$ seconds and its $155{,}929$ exhaustive decisions take $86.8$ seconds across $24$ CPU workers, with zero unknowns. These audit timings exclude the final statistical calculation.

For Theorem 4.1, the implemented rule's effective predictor is $q - \eta$ outside the last-card boundary and $B = -1$ at that boundary. It therefore stops exactly when this effective predictor is at least the cap; the residual in the theorem is computed for this effective predictor.

**Paired inference.** The eight fixed endpoints are the original five rules, first denial, anchored continuation and Monte Carlo prediction. Their supports require $84 + 3 \cdot 21 = 147$ integer tails. We also fix three paired contrasts: anchored minus last-card denial, Monte Carlo minus last-card denial, and anchored minus Monte Carlo. Both learned rules preserve the last-card score on every game where that rule triggers, verified on all $9{,}894$ such evaluation games. Each difference from last-card denial therefore lies in $[-7, 14]$. The mutual difference lies in $[-14, 14]$, since two triggers on the same ordinary game have the same score $U + 1$. These contrasts add $21 + 21 + 28 = 70$ tails. All reported lower limits for this study allocate $0.05/217$ to each one-sided exact binomial tail, simultaneously covering the eight endpoints and three contrasts. Standard errors use the unbiased variance of complete-game scores; paired errors use the variance of their gamewise differences. Model and threshold selection uses only the separate development sample.

| Rule | Trigger rate | Mean score | SE | 95% lower |
|---|---|---|---|---|
| First true-deal throw | 0.03200 | 0.03137 | 0.00291 | 0.01586 |
| First information-chargeable throw | 0.01572 | 0.00743 | 0.00160 | −0.00124 |
| Last-card true-deal throw | 0.01008 | 0.01190 | 0.00129 | 0.00748 |
| Last-card information-chargeable throw | 0.00962 | 0.01023 | 0.00119 | 0.00620 |
| Last-card denial | 0.16490 | 0.31042 | 0.00670 | 0.28459 |
| First denial | 0.63188 | −0.73982 | 0.01200 | −0.80579 |
| Anchored continuation | 0.20962 | 0.40180 | 0.00780 | 0.36702 |
| Monte Carlo prediction | 0.21173 | 0.41103 | 0.00789 | 0.37565 |

Table 8: All eight frozen rules on the independent 60,000-game DMC cohort. Scores are points per game. Lower limits are simultaneous over these endpoints and the three paired contrasts, using the 217-tail allocation. Every negative score and zero-trigger game contributes to its estimator.

### F.1 Executed deterministic continuation and residuals

The deterministic oracle of Appendix A.7 is an analytically specified follow-up on this same 60,000-game evaluation cohort; both fitted predictors and their thresholds remain frozen. We replay every game and copy its complete state after the first certified denial, including the public last-sixteen-action history consumed by DMC's LSTM. The LSTM recomputes from that history without a persistent hidden state; evaluation-mode argmax and game transitions are deterministic. The coalition's hands and public played cards determine the remaining focal hand. Forecasting receives only the copied current state and frozen policy, with empty reward fields. All 37,913 forecasts exactly reproduce the ordinary terminal payoff, and all 60,000 replayed games reproduce their initial deal and payoff, with zero illegal actions or unfinished games.

The oracle switches at that first denial exactly when the forecast exceeds $-1$. Table 2 gives its cap-valued optimum and paired regrets. The seven endpoints require 112 integer thresholds; allocating $0.05/224$ to each threshold and direction gives simultaneous two-sided exact intervals. State reconstruction takes 16.29 seconds; continuation forecasting takes 0.94 seconds for 458,142 active GPU decisions, using the loaded policy and captured states. These costs are separate from the original 86.80-second certificate search.

For execution, the first 512 denial states in game order form an outcome-independent witness prefix. Opponents execute exact denial witnesses while the focal player follows DMC with its evolving public history. All 512 terminate with focal payoff at most $-1$, with zero focal wins, illegal actions or fallbacks. The oracle selects 56 of these states; its mean cap-valued and actual gains over this conditional prefix are 0.83594 and 0.86523. Execution takes 1.51 seconds, including 1.30 seconds of witness search. Replay verifies every transition and all 1,238 focal DMC actions. The prefix measures actual execution conditional on a denial; Table 2 uses all complete games.

At the 125,743 ordered denial opportunities, deterministic continuation makes the observed next value equal its conditional expectation. We evaluate the effective predictors defined above and their true residuals on this subsequence, extending the initial transition with zero residual. The global quantities $R(q)$ and $R_{\pm}(q)$ are defined in Appendix A.3. For the oracle-dependent diagnostics, define

$$R_{\mathrm{win}}(q) := \mathbb{E}_{\pi} \sum_{t=0}^{H-1} \mathbf{1}\{\min(\tau_q, \tau^*) \le t < \max(\tau_q, \tau^*)\} |\delta_t|, \qquad Q(q) := \mathbb{E}_{\pi} \sum_{t=0}^{H-1} |q_t - q_t^*|,$$

where $\tau^*$ and $q^*$ are the optimal stopping time and continuation function. Table 9 reports all four quantities; every gamewise regret bound holds. Evaluation takes 14.41 seconds.

| Predictor | Regret | $R$ | $R_{\pm}$ | $R_{\mathrm{win}}$ | $Q$ |
|---|---|---|---|---|---|
| Anchored | 0.11025 | 1.30796 | 0.48088 | 0.20192 | 1.63982 |
| Monte Carlo | 0.10102 | 1.47352 | 0.73778 | 0.28578 | 1.79770 |

Table 9: True conditional-residual diagnostics under full deterministic DMC play. The global and one-sided bounds use the supplied predictor; the stopping-window and continuation-error diagnostics additionally use the optimum.

### F.2 MATCHED PREDICTION CAPACITY

A further comparison fixes every deployed predictor at 64 trees and retains the same 125 features, tree parameters, clipping, last-card anchor and validation margin grid. Five independently randomized $48{,}000/12{,}000$ partitions reuse the development cohort. Anchored return regression and Monte Carlo regression each fit once. Fitted continuation and realized-return policy iteration (the cash-flow regression comparator) each fit eight stages, sharing the first anchored fit. The latter uses target $-1$ if the previous zero-margin rule stops at a strictly later denial, and $U$ otherwise. Validation selects stage and margin with the same tie rule as above. The 80 unique fitted predictors are frozen before scoring a new $60{,}000$-game cohort. Generation runs concurrently with fitting; fitting accesses only development games, and evaluation verifies frozen models before reading test data. All assigned games complete, with no initial-deal overlap with either earlier cohort.

| Selected 64-tree rule | Mean | Game SE | Partition SE | Fit seconds |
|---|---|---|---|---|
| Anchored return | 0.38923 | 0.00760 | 0.00060 | 3.21 |
| Monte Carlo | 0.38692 | 0.00768 | 0.00075 | 3.14 |
| Fitted continuation | 0.38832 | 0.00757 | 0.00116 | 25.98 |
| Realized-return iteration | 0.38772 | 0.00753 | 0.00049 | 25.46 |

Table 10: Matched deployed capacity on the further independent cohort. Means first average the five frozen rules within each game. Partition SE describes five overlapping fitting partitions of the same development population. Fit time averages all stages per partition.

The last-card score is 0.30347 (game SE 0.00664), first-denial score is $-0.76628$ (SE 0.01197), and the analytic constant-cap optimum is 0.49980 (SE 0.00815). Anchored regression improves on last-card denial by 0.08576 with paired game interval $[0.07827, 0.09326]$. The original frozen anchored and 512-tree Monte Carlo predictors score 0.39168 and 0.39857 on this cohort. Fitted continuation selects stages $(1, 1, 5, 1, 1)$ and realized-return iteration selects $(1, 8, 5, 6, 8)$ across the five partitions.

| Paired contrast | Mean | 95% game interval | 95% partition interval |
|---|---|---|---|
| Continuation − anchored | $-0.00091$ | $[-0.00156, -0.00027]$ | $[-0.00345, 0.00162]$ |
| Continuation − Monte Carlo | 0.00140 | $[-0.00053, 0.00333]$ | $[-0.00142, 0.00422]$ |
| Return iteration − anchored | $-0.00151$ | $[-0.00311, 0.00010]$ | $[-0.00317, 0.00016]$ |
| Return iteration − continuation | $-0.00059$ | $[-0.00190, 0.00071]$ | $[-0.00416, 0.00298]$ |
| Anchored − Monte Carlo | 0.00231 | $[0.00049, 0.00413]$ | $[-0.00037, 0.00499]$ |

Table 11: Paired capacity controls. Game intervals use $\bar{\Delta} \pm 1.96\,\mathrm{SE}$ for the within-game five-policy average; partition intervals use $t_{4,0.975}$ across the five paired partition effects, conditional on the shared development and test populations. These are marginal intervals, separate from the exact simultaneous families above.

| Stage | Fitted continuation mean ± partition SE | Return iteration mean ± partition SE |
|---|---|---|
| 1 | $0.38923 \pm 0.00060$ | $0.38923 \pm 0.00060$ |
| 2 | $0.38471 \pm 0.00039$ | $0.38736 \pm 0.00052$ |
| 3 | $0.38300 \pm 0.00092$ | $0.38630 \pm 0.00068$ |
| 4 | $0.38376 \pm 0.00068$ | $0.38881 \pm 0.00047$ |
| 5 | $0.38391 \pm 0.00023$ | $0.38755 \pm 0.00078$ |
| 6 | $0.38359 \pm 0.00054$ | $0.38850 \pm 0.00056$ |
| 7 | $0.38227 \pm 0.00099$ | $0.38756 \pm 0.00087$ |
| 8 | $0.38287 \pm 0.00041$ | $0.38789 \pm 0.00062$ |

Table 12: Every fitting stage, with each margin selected on validation games. Test outcomes do not select a stage.

All signed and zero scores enter these comparisons. Fitting and validation take 315.7 seconds on two CPU threads; the new audit takes 17.65 seconds for GPU rollout and 126.75 seconds for exhaustive search across 24 CPU workers. Predicting with all 80 models and the two contextual predictors takes 37.31 seconds. Equal deployed capacity and the additional work used to explore stages are reported separately.

### F.3 COMPLETE GAMES WITH A SAMPLED POLICY

**Information and population.** The frozen PPO control policy has $507$ observation inputs and $11.72$ million stored parameters. Every seat samples its native masked categorical policy at temperature one. We use three independently generated development cohorts, each with $48{,}000$ fit and $12{,}000$ validation games, and one independent $60{,}000$-game test cohort. All models and forecast margins are frozen before test generation. All $240{,}000$ assigned games complete, with no initial-deal overlap between any two cohorts. At the $844{,}577$ candidate decisions, mean action entropy is $0.09051$ nats and $25.86\%$ have entropy above $10^{-3}$ nats. The test cohort contains $211{,}338$ resolved candidate decisions, of which $171{,}549$ are certified denials. No decision is unknown and no game is excluded.

The $593$ predictor features extend the earlier $125$ features with four $52$-bit public played-card masks, four $52$-bit last-move masks and the $52$-bit actual table. They include the current hands, counts and observed focal action. These features reconstruct exactly every state field used by this feed-forward policy, legal-action generation and subsequent transitions. The coalition can obtain them from its joint observations. The action is applied once before forecasting. Public-history and episode bookkeeping fields retained during collection do not affect this policy or these transitions. Original terminal outcomes and future random draws are absent from forecast inputs.

**Matched fitting and selection.** Every deployed regressor has $64$ histogram-boosted trees, learning rate $0.08$, at most $31$ leaves per tree, minimum leaf population $80$, $\ell_2$ regularization $1$, and $63$ histogram bins; early stopping is disabled. Predictions are clipped to $[-8, 13]$. Anchored regression fits the realized return of continuing until a future last-card denial, where the target becomes $-1$. Outcome regression fits the ordinary terminal payoff. Fitted continuation and cash-flow iteration use the eight-stage targets of Appendix F.2, sharing the first anchored fit. There are $48$ unique fits across the three development populations. The same margin grid $\{0, 0.1, 0.25, 0.5, 1\}$ and last-card anchor apply to all methods; separate validation games maximize the signed score, breaking ties by smaller stage and then smaller margin.

The ordinary suffix controls use $8$ and $32$ independent sampled continuations, with the first eight shared between budgets. Each branch starts from the observed current state, uses fresh action randomness independent of actual subsequent play, and completes the game. The exact last-card anchor bypasses forecasting. Only the first development cohort's fixed $12{,}000$ validation games select forecast margins; both select $1$. The forecasts are frozen once and are shared across the three learned-rule comparisons. At $32$ branches, $26.51\%$ of nonlast queries have nonzero terminal-return sample variance, whose mean is $0.7201$.

| Learned rule | Mean | Development SE | Fit seconds | Prediction seconds |
|---|---|---|---|---|
| Anchored return | 0.34932 | 0.00100 | 7.92 | 0.419 |
| Outcome regression | 0.34623 | 0.00237 | 6.30 | 0.416 |
| Fitted continuation | 0.34932 | 0.00100 | 51.99 | 0.419 |
| Cash-flow regression | 0.35176 | 0.00290 | 51.18 | 0.439 |

Table 13: Independent-development variation and component cost. Development SE uses three disjoint training/validation populations conditional on the common test cohort. Fit time includes every candidate stage and averages development cohorts; shared first fits contribute to each method's standalone cost. Prediction time is a measured batch of required queries through each first stop, using eight CPU threads.

For each learned method, the reported test score averages its three frozen gates within each game. This is the value of selecting a developed gate uniformly at the start of a game. Game SE uses this average once per independent test game. Development intervals instead use paired $t_{2,0.975}$ intervals over the three independent development effects, conditional on the shared test population. They measure separate sources of variation.

The exact simultaneous lower bounds in Table 3 allocate $0.05/735$ to each support-threshold limit across seven endpoints and these six contrasts. Individual earlier-switch scores have support $[-7, 14]$, last-card scores $[0, 14]$, and paired constant-cap differences $[-14, 14]$. Learned averages are on a grid of thirds. The last-card rule contributes $14$ thresholds, the four learned-rule averages $4 \cdot 63$, and the two forecasts $2 \cdot 21$. The six contrasts contribute $63 + 4 \cdot 84 + 28$ thresholds, giving $308 + 427 = 735$ in total.

| Paired contrast | Mean | 95% game interval | 95% development interval |
|---|---|---|---|
| Anchored − last-card | 0.21375 | [0.20156, 0.22594] | [0.20945, 0.21805] |
| Continuation − anchored | 0 | [0, 0] | [0, 0] |
| Continuation − outcome | 0.00309 | [0.00072, 0.00546] | [−0.01091, 0.01709] |
| Cash-flow − anchored | 0.00244 | [0.00084, 0.00404] | [−0.00805, 0.01293] |
| Anchored − forecast 32 | −0.19498 | [−0.20304, −0.18693] | [−0.19928, −0.19069] |
| Forecast 32 − 8 | 0.00247 | [0.00098, 0.00395] | — |

Table 14: The six fixed contrasts on the same complete test games. Game intervals use 1.96 SE and are marginal, conditional on frozen rules. The forecast-only contrast has one validation-selected rule per budget. The zero continuation contrast follows from selecting the identical first fit in all three development cohorts.

The anchored-minus-last-card support is $[-7, 14]$: every positive last-card game has the same payoff-cap score under an anchored earlier switch. All negative and zero terms enter the estimates. The simultaneous lower limit for anchored-minus-last-card is 0.18035; small differences between learned methods and between forecast budgets have nonpositive simultaneous lower limits.

| Stage | Continuation mean ± development SE | Cash-flow mean ± development SE |
|---|---|---|
| 1 | $0.34932 \pm 0.00100$ | $0.34932 \pm 0.00100$ |
| 2 | $0.34603 \pm 0.00195$ | $0.34644 \pm 0.00250$ |
| 3 | $0.34302 \pm 0.00240$ | $0.35187 \pm 0.00214$ |
| 4 | $0.34177 \pm 0.00227$ | $0.34898 \pm 0.00245$ |
| 5 | $0.34089 \pm 0.00298$ | $0.35096 \pm 0.00314$ |
| 6 | $0.34134 \pm 0.00179$ | $0.34886 \pm 0.00133$ |
| 7 | $0.34203 \pm 0.00291$ | $0.35068 \pm 0.00170$ |
| 8 | $0.34168 \pm 0.00194$ | $0.35063 \pm 0.00206$ |

Table 15: Every fitted stage, using its validation-selected margin. Test scores do not select a stage. Cash-flow selection chooses the first fit in two development cohorts and the fifth in the third; continuation selects the first in all three.

**Measured work.** Test ordinary rollout takes 26.65 seconds. Exact labeling takes 40.41 seconds on 24 CPU workers; labeling, data preparation and feature reconstruction together take 61.45 seconds. Fitting and validation take 114.15, 111.29 and 107.42 seconds across development cohorts. These costs precede or are shared by the switching comparisons.

Required-query forecasting exposes one current denial per still-active game in each scheduling round and stops querying that game after its first switch. The 8-branch rule uses 134,655 non-last queries, 1,077,240 branches and 11,864,778 suffix decisions; the 32-branch rule uses 135,122 queries, 4,323,904 branches and 47,633,339 suffix decisions. Forecast time is $13.69/50.34$ seconds on one RTX 5090; including query scheduling gives $13.84/50.57$ seconds. Current-state reconstruction adds $4.15/4.30$ seconds and loading adds $0.59/0.62$ seconds. The corresponding all-query forecasts cost $16.29/61.36$ seconds. These are measured component times: the CPU predictor timing batches required rows, whereas the GPU forecast timing includes sequential query rounds. Ordinary play and common certification are accounted for separately.

Every required-query forecast and active-decision count matches the all-query computation exactly. Changing global random-generator state, branch order and nonfunctional bookkeeping leaves the independently keyed forecasts unchanged. All branches finish legally. Independent scalar replay verifies every selected rule's complete-game score; 179 checks cover chronology, information reconstruction, cohort separation and statistical recomputation with zero failures.

**Executing the sampled-policy refutation.** We take the first denial from each of the first 512 eligible test games, without inspecting ordinary terminal outcomes. Opponents execute exact witnesses while the focal player follows the frozen sampled policy with fresh randomness. All 512 settle at focal payoff at most $-1$, with no focal wins, illegal actions, fallbacks or unfinished games. Replay matches every transition and all 1,511 sampled focal actions. Execution takes 1.47 seconds, including 0.83 seconds of witness search; state reconstruction and policy loading add 2.52 and 0.62 seconds. This check measures cap compliance and execution work on the fixed conditional prefix.

### F.4 NON-ENUMERATIVE STOCHASTIC FULL-GAME CERTIFICATION

**Fixed rules and population.** We use the first $32{,}768$ games of the independently completed sampled-PPO test cohort, fixing this prefix from compute requirements before evaluating the new dual. Games without eligible denials contribute zero. The continuation predictor is the first development cohort's frozen anchored 64-tree model, with validation-selected margin $0.5$. The second rule is the already frozen 32-branch ordinary forecast with margin $1$. Both retain the exact last-card anchor. Their achieved cap gains on this prefix are $0.34628$ (game SE $0.01001$) and $0.54565$ (SE $0.01134$). The paired forecast-minus-learned gain is $0.19937$ (SE $0.00625$).

The previously completed ordinary cohort and frozen rules are reused without retuning. Prefix size comes from compute estimates; the clamp, absorption and support bounds below come from physical game rules. These choices and both inner budgets are fixed before evaluating the new dual scores.

For a denial leaving $n$ focal cards, let $q$ be the raw output of the frozen continuation predictor, and write $\mathrm{clip}(x, a, b) := \min\{b, \max\{a, x\}\}$. We set

$$q_{\text{eff}} = \max\{-n, \mathrm{clip}(q, -8, 13) - 0.5\}, \qquad v = \min\{-1, q_{\text{eff}}\},$$

with $q_{\text{eff}} = v = -1$ at last-card denials. This transformation is fixed before evaluation. At nonlast denials, $-n \le -2$, so the lower clamp leaves every selected switch unchanged. It also ensures $v \ge -n$, a physical lower bound on subsequent ordinary payoff. Proposition A.3 terminates the certificate process at the first last-card denial. The original complete-game payoff remains the reference for achieved gain.

**Independent next-eligible simulation.** From each eligible post-action state before absorption, we draw $32$ fresh ordinary suffixes; the first eight also supply the smaller budget. Each uses the frozen PPO policy at its native sampling temperature. Exact search scans subsequent focal decisions to find the first eligible denial. The target is the frozen $v$ there, or the ordinary terminal payoff if no denial occurs. All suffixes complete for validation, but outcomes after an eligible successor do not enter its target. Current-state queries contain no original terminal payoffs. Auxiliary draws use separate random streams from both original play and the forecast gate's simulations.

The martingale is zero before the first eligible history and accumulates $v_{s+1} - \widehat{m}_s$ at compressed eligible transitions. We evaluate $Z_{\rho,q,k}$ for both fixed rules $\rho$, using the same learned martingale and absorbed boundary. Every score is nonnegative; optional sampling gives $F^* - F(\rho) \le \mathbb{E} Z_{\rho,q,k}$. This expectation includes finite inner-simulation noise.

**Analytical support.** Immediately after an eligible action, $n \le 8$ focal cards and $p \le 13$ opponent cards remain; a pass can retain all eight focal cards. Write $r = n + p \le 21$. Between consecutive eligible focal post-action histories, at least one card leaves a hand: an entire rotation of passes cannot continue through a free lead. Thus $r$ strictly decreases, and each nonterminal state has $r \ge 4$. Every successor target lies in $[-n, p]$, since nonterminal values obey the retained-card lower bound and a focal win pays at most the opponent-card total.

For any chosen stop $\rho$, a candidate $t < \rho$ contributes

$$(B_\rho - v_\rho) - (B_t - v_t) + \sum_{s=t}^{\rho-1} (\widehat{m}_s - v_s)$$

to the paired maximum. Let $r_s$ denote the value of $r$ at compressed transition $s$. For $t > \rho$, reverse the summand's sign and sum from $\rho$ to $t-1$. The first sum is at most $\sum_s r_s \le 21+20+\cdots+4 = 225$; the reversed sum is at most $7 \cdot 18 = 126$. Terminal boundary gaps vanish and other boundary gaps are nonnegative. The greedy rule has $B_{\tau_q} - v_{\tau_q} = 0$, so $0 \le Z_{\tau_q,q,k} \le 225$. The forecast's chosen-boundary gap is at most 7, giving $0 \le Z_{\rho,q,k} \le 232$. These bounds use game rules and the frozen transformation, with no observed-maximum substitution.

**Finite-sample inference.** Let $n_{\text{game}}$ be the number of independent complete games, let $\bar{Z}$ and $S^2$ be the sample mean and unbiased sample variance of $Z_{\rho,q,k}$, and let $L$ be its support upper bound. For a failure probability $\alpha \in (0, 1)$, applying Maurer & Pontil (2009, Theorem 4) to $Z_{\rho,q,k}/L$ gives

$$F^* - F(\rho) \le \bar{Z} + \sqrt{\frac{2S^2 \log(2/\alpha)}{n_{\text{game}}}} + \frac{7L \log(2/\alpha)}{3(n_{\text{game}} - 1)}$$

with probability at least $1 - \alpha$. We use $\alpha = 0.05/4$ for the two fixed rules and two fixed inner budgets. The union bound gives simultaneous $95\%$ coverage for all four regret upper bounds despite shared games and successor samples.

After evaluation, this four-endpoint allocation combines the two initially separate two-budget confidence families. It conservatively widens their intervals without changing any rule, cohort, support or score.

| Paired rule | Inner $k$ | Mean $Z$ | Outer SE | Joint $95\%$ regret upper |
|---|---|---|---|---|
| Anchored continuation | 8 | 0.210149 | 0.005933 | 0.310366 |
| | 32 | 0.209501 | 0.005909 | 0.309644 |
| Ordinary forecast, 32 branches | 8 | 0.013753 | 0.000743 | 0.099965 |
| | 32 | 0.012469 | 0.000663 | 0.098429 |

Table 16: Independent-successor duals on 32,768 complete ordinary sampled-policy games. Upper endpoints are rounded outward. The forecast's 32 action-simulation branches are distinct from the dual's inner budget $k$.

For the forecast at $k = 32$, the upper endpoint decomposes as $0.012469$ (mean) $+ 0.002113$ (variance term) $+ 0.083845$ (support term). The last term depends on the outer-game count; the inner budget affects the first two terms.

At $k = 32$, the learned rule's sampled one-sided residual envelope averages $0.75843$, while its paired dual averages $0.20950$. The same martingale paired with the forecast gives $0.01247$. As a separate centering diagnostic, the forecast's paired score has SE $0.000663$, compared with $0.001484$ for $Z_{\rho,q,32} + \widehat{M}_\rho$. These same-path estimators have the same expectation by optional sampling; subtracting the chosen-stop martingale reduces measured variance by a factor of $5.0$ in this cohort.

**Cost and checks.** There are 78,867 current-state queries, 2,523,744 independent suffixes and 28,504,633 ordinary branch decisions. Exact next-eligible adjudication checks 2,189,204 candidates, with no unknowns. All branches complete with legal moves. GPU play, capture and serialization take $62.61$ seconds; CPU adjudication and feature prediction take $334.17$ seconds across $24$ processes on the shared machine. Ordinary-game collection, model fitting and the frozen forecast gate have their separately reported costs. Additional dual evaluation takes $396.78$ seconds after those inputs are available.

Every one of the 32,768 original learned-gate scores and forecast scores is reproduced exactly. Checks cover current-state reconstruction, absence of future labels, global-random-state independence, exact repeat and branch-order invariance. Changing branch streams changes sampled actions. A disjoint correctness prefix verifies first-eligible targets against the established full-game audit. Among current queries, 20,017 have positive inner-target variance. Analytical support, nonnegativity, score recomputation and the confidence formula pass on every evaluated game.

## G Training details and complete controls

Both real-game trainers use fused proximal policy optimization (PPO) with discount factor $\gamma = 1$, generalized advantage estimation (GAE) parameter $0.99$, learning rate $10^{-4}$, AdamW coefficients $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay $10^{-4}$, gradient clipping at 1, and rollout segments of 32 decisions. Big Two uses $8{,}192$ lanes and $11{,}444$ updates ($2{,}999{,}975{,}936$ decisions); TLMN uses $4{,}096$ lanes and $5{,}722$ updates ($749{,}993{,}984$ decisions). Certified resets use $\zeta = 0.08$ and auxiliary weight $\lambda = 0.3$. All arms use their game's fixed architecture and final checkpoint. Each comparison uses five independent paired runs.

Both policy networks use six residual feed-forward trunk layers of width $512$, with feed-forward widths $1{,}536$ (Big Two) and $1{,}024$ (TLMN), and two value layers. Big Two has a masked categorical action head; TLMN has a hierarchical card-selection head. TLMN's value head receives the complete hands during training in both arms; its policy head observes only the focal information state. The certificate labels likewise depend only on that information state.

For $m = |A^{\mathrm{pool}}(I)|$, define the policy conditioned on the certified set by $\bar{\pi}(a \mid I) := \pi(a \mid I)/\pi(A^{\mathrm{pool}}(I) \mid I)$ for $a \in A^{\mathrm{pool}}(I)$. The TLMN surrogate satisfies the exact identity

$$\ell_{\mathrm{CE}} = \ell_{\mathrm{cert}} + \log m + \mathrm{KL}\big(\mathrm{Unif}(A^{\mathrm{pool}}(I)) \,\|\, \bar{\pi}(\cdot \mid I)\big),$$

where KL denotes Kullback–Leibler divergence. This follows by substituting $\pi(a \mid I) = \pi(A^{\mathrm{pool}}(I) \mid I)\bar{\pi}(a \mid I)$ into the uniform cross entropy. Both games supervise the same kind of certified set; TLMN's sampled cross entropy also regularizes the within-set distribution. The comparison between games is reported at their implemented objectives.

| Game/init | Arm | Greedy $\rho_{\mathrm{deal}}$ | Payoff [training 95% CI] | Game SE |
|---|---|---|---|---|
| Big Two/scratch | Control | 0.97712 | — | — |
| | Certified resets + aux | 0.97784 | $-0.02808$ $[-0.04261, -0.01355]$ | 0.01468 |
| | Certified resets only | 0.97735 | $-0.01604$ $[-0.02499, -0.00710]$ | 0.01471 |
| | Uncertified resets | 0.97716 | $-0.02752$ $[-0.04823, -0.00681]$ | 0.01471 |
| Big Two/clone | Plain fine-tune | 0.97984 | — | — |
| | Certified fine-tune | 0.98211 | $-0.05256$ $[-0.09298, -0.01215]$ | 0.01486 |

Table 17: All Big Two five-run campaign controls. Retention uses the same 5,289 states; the initialization clone has retention 0.83021. Each payoff row completes 200,000 games per pair. Training $t_4$ intervals condition on the shared cohort; game SE conditions on the five frozen pairs and retains their within-game covariance.

The corresponding complete-cohort TLMN certified-versus-control payoff is $+0.00110$ stakes per game, with training 95% interval $[-0.00250, 0.00469]$ and game SE 0.00244. Marginal normal game intervals for the four Big Two payoff rows are $[-0.05685, 0.00070]$, $[-0.04488, 0.01280]$, $[-0.05635, 0.00131]$ and $[-0.08168, -0.02345]$; TLMN gives $[-0.00369, 0.00589]$. These quantify evaluation-game variation separately from the training intervals in Table 17.

All 25 matchups complete their assigned 5,000,000 games with no illegal action or unfinished game. Independent initial-cohort reconstruction verifies every deal digest and all terminal-at-deal TLMN outcomes; score recomputation verifies reward units, checkpoint assignments and both uncertainty calculations. Summed rollout time is 319.03 seconds, excluding loading, output serialization and independent verification.

| TLMN natural | Metric | Control | Certified | Paired change [95% CI] |
|---|---|---|---|---|
| Greedy | $\rho_{\mathrm{deal}}$ | 0.99711 | 0.99703 | $-0.00008$ $[-0.00113, 0.00097]$ |
| Greedy | $\rho_{\mathrm{pool}}$ | 0.99372 | 0.99298 | $-0.00074$ $[-0.00274, 0.00126]$ |
| Sampled | $\rho_{\mathrm{deal}}$ | 0.99517 | 0.99518 | $+0.00001$ $[-0.00107, 0.00108]$ |
| Sampled | $\rho_{\mathrm{pool}}$ | 0.98986 | 0.98997 | $+0.00011$ $[-0.00163, 0.00185]$ |

Table 18: The same five pairs of TLMN final checkpoints on the separate natural held-out population (2,422 states). Sampling and interval conventions match Table 5.

| Game | Decisions (B) | Control (h) | Certified (h) | Ratio | Pool build (s) |
|---|---|---|---|---|---|
| Big Two | 3.000 | 36.53 | 38.38 | 1.051 | 649 |
| TLMN | 0.750 | 36.32 | 67.06 | 1.846 | 401 |

Table 19: Measured five-run average run times at matched decisions on one RTX 5090. Run time includes compilation and in-loop evaluation; process setup before training instrumentation is excluded. Pool construction is a separate shared preprocessing cost, reused across the five independent runs.

Post-compilation loop times are $35.25/37.10$ hours for Big Two control/certified and $31.02/57.67$ for TLMN. Big Two's 649-second preprocessing comprises 119 seconds of harvest and 530 of labeling. Amortizing the TLMN pool across its five certified runs adds 80.2 seconds per run. The decision budget, measured run time and preprocessing cost remain separate quantities.

# H SMALL-GAME VALIDATION AND CERTIFICATE COVERAGE

**Exact stopping evaluation.** We sample 32 Big Two v2 endgames with two to four focal cards and two or three cards in each opponent's hand; the focal player leads an empty table. All four ordinary policies choose uniformly among legal actions. Complete enumeration covers 200,716 states across these endgames and computes ordinary values, the optimal cap-valued stopping rule, and the payoff of an extracted refutation. Eligible boundaries occur after a focal action with a proved losing continuation and use cap $B = -1$.

For each endgame, 4,000 complete ordinary trajectories estimate transition frequencies. Backward induction on this empirical transition model fits one continuation function; an unobserved state uses $q = -n_{\text{focal}}$, the negative number of focal cards remaining. This implements the last-card rule at unobserved denials. Complete trajectories ensure that every observed nonterminal successor has outgoing transition counts, so unobserved fallback values do not enter observed-state backups. Exhaustive checks verify the stopping decisions and fitted backups. Fitting uses 128,000 ordinary trajectories. Exact enumeration separately evaluates the fitted rule and its refutation over all reachable states. The resulting cap-valued floor averages 0.75707, against the exact optimum 0.76230, a mean gap of 0.00523; its global residual is 0.56213, one-sided residual 0.27955, stopping-window residual 0.01827, and continuation error 1.39912.

To test Theorem 4.1, we perturb the exact continuation value independently at each state by uniform noise of magnitude 0.05, 0.2, 1 or 5, with five predictors at each magnitude per endgame. Each of the 640 checks evaluates both the policy gap and the sum of absolute conditional Bellman residuals by exact enumeration under ordinary play. All satisfy the coefficient-one bound. We also verify the optimal stopping identity, the payoff-cap inequality for the extracted refuter, and the gamewise score identity on every training trajectory. These checks use exact conditional expectations, while the larger learned study attaches confidence directly to independent complete-game scores.

## H.1 EXACT LEARNING CURVES

For each of the same 32 endgames, five new independent ordinary-trajectory streams give nested training prefixes of 250, 1,000, 4,000 and 16,000 trajectories. The fitting rule and unobserved-state convention remain fixed across all budgets. We average the five repetitions within each endgame, then average the 32 endgames. Table 20 reports exact conditional residuals and policy values, with the global norm and both refinements defined above. Each fitted rule satisfies all displayed regret bounds; exact refutation gains also exceed their cap-valued scores. The analytic last-card rule has mean gain 0.69855.

| Trajectories | Cap gain | Actual gain | Regret | $R$ | $R_{\pm}$ | $R_{\text{win}}$ | $Q$ |
|---|---|---|---|---|---|---|---|
| 250 | 0.73129 | 0.81016 | 0.03100 | 1.60299 | 0.76742 | 0.11482 | 4.74102 |
| 1,000 | 0.74886 | 0.81891 | 0.01344 | 0.98003 | 0.48151 | 0.04207 | 2.67287 |
| 4,000 | 0.75790 | 0.82624 | 0.00440 | 0.55860 | 0.27957 | 0.01623 | 1.38346 |
| 16,000 | 0.76094 | 0.82702 | 0.00135 | 0.28959 | 0.14600 | 0.00442 | 0.68391 |

Table 20: Learning on ordinary trajectories with exact evaluation in 32 stochastic Big Two endgames. Regret is relative to the cap-valued optimum 0.76230; actual gain evaluates the extracted refutation. $R_{\text{win}}$ and $Q$ use the exact optimal oracle.

The 2.56 million ordinary training trajectories take 11.09 seconds with eight CPU workers; replaying the original 128,000-trajectory fit takes 1.07 seconds. Enumeration of refutation suffixes is used only for evaluation. Increasing the budget reduces the median endgame regret from 0.01535 to 0.00032 and the maximum from 0.11421 to 0.01067. The residual and continuation-error norms are both evaluated directly; their relation depends on the fitted predictor, with $Q < R$ in five of the 640 fits across the four budgets.

## H.2 STOCHASTIC ENDGAME DUAL CALIBRATION

**Design and exact evaluation.** We use 32 original and 32 independently generated Big Two endgames, with 2–4 focal cards and 2–3 cards per opponent. New roots are generated without

filtering by outcomes. Ordinary play chooses uniformly among legal actions. Five independent fitting streams per root use nested budgets of 250, 1,000, 4,000, and 16,000 complete ordinary paths. Empirical transition backups fit the continuation function; an unseen state uses minus the number of focal cards as its bounded fallback. The candidate family therefore contains 1,280 fixed predictors. Repetitions are averaged within each root before averaging roots.

The original and new cohorts contain respectively 200,716 and 120,161 ordinary reachable nonterminal states. Their trees contain 608,625 and 168,379 complete paths in total; the largest individual tree has 277,951 paths. Exhaustive ordinary-path evaluation computes both $\mathcal{D}(q) = \mathbb{E}Z_q$ and $\max Z_q$ for every fixed predictor. All 1,280 checks satisfy $F^* - F(\tau_q) \le \mathcal{D}(q) \le R_{\pm}(q) \le R(q)$ within $10^{-8}$. The cohort results are:

| | Original 32 roots | | New 32 roots | |
|---|---|---|---|---|
| Training paths | Regret | $\mathcal{D}$ | Regret | $\mathcal{D}$ |
| 250 | 0.031005 | 0.134042 | 0.014330 | 0.087713 |
| 1,000 | 0.013441 | 0.075375 | 0.007416 | 0.043894 |
| 4,000 | 0.004398 | 0.036066 | 0.002362 | 0.018006 |
| 16,000 | 0.001354 | 0.013411 | 0.000825 | 0.006647 |

These regrets compare certified stopping bounds. Actual refutation payoffs are evaluated separately from their caps.

**Monte Carlo calibration.** Before exhaustive evaluation, each fixed candidate is scored on 32,768 independent ordinary test paths. Let $\overline{Z}_q$ be the sample mean of these scores and let $L_q = \max Z_q$, as verified by full path enumeration. With $n = 32{,}768$ and $\alpha = 0.05/1{,}280$, Hoeffding's inequality gives simultaneously

$$\mathcal{D}(q) \le \min\left\{ R_{\pm}(q), L_q, \overline{Z}_q + L_q\sqrt{\frac{\log(1/\alpha)}{2n}} \right\}$$

with probability at least $0.95$, conditional on the fitted family. The same test samples and familywise allocation are retained when the range is refined from the maximum one-sided path sum to $\max Z_q$. The following entries average candidate-specific upper bounds, first within roots and then across all 64 roots:

| Training paths | MC mean $\overline{Z}$ | Initial upper | Refined upper |
|---|---|---|---|
| 250 | 0.110942 | 0.252067 | 0.199717 |
| 1,000 | 0.059809 | 0.195399 | 0.144896 |
| 4,000 | 0.027117 | 0.149056 | 0.103095 |
| 16,000 | 0.010115 | 0.102315 | 0.071415 |

The learning-curve table uses exhaustive expectations; this separate calibration quantifies uncertainty from sampled outer paths.

**Independent successor simulation.** We also replace each exact conditional mean by the mean of $k$ independent auxiliary successors, using the first 4,000-path fit at each root and 8,192 outer paths. The classical nested dual construction (Andersen & Broadie, 2004) produces the following paired estimates (mean $\pm$ aggregate outer Monte Carlo standard error):

| Inner samples $k$ | Fitted predictor | Optimal-predictor control |
|---|---|---|
| 1 | $0.294139 \pm 0.001375$ | $0.275917 \pm 0.001247$ |
| 4 | $0.104698 \pm 0.000684$ | $0.079698 \pm 0.000513$ |
| 16 | $0.046498 \pm 0.000436$ | $0.018992 \pm 0.000179$ |

Their exact-conditional dual values are respectively $0.027070$ and zero. The optimal-predictor control has $Z_{q^*} = 0$ on every outer path with exact conditional means. Finite inner simulation adds a noise floor while preserving an expected upper bound; its envelope uses the noisy conditional residual, rather than the exact-conditional residual.

**Frozen selection criteria.** Within each fitting stream, minimum validation dual score and maximum validation signed payoff score select from the same four budgets using 8,192 common validation paths. Choices are frozen before test evaluation; neither criterion uses $q^*$, exact regret, or test dual scores. The dual criterion additionally evaluates the ordinary conditional operator. The pooled exact regrets are $0.0010867$ for dual selection, $0.0012036$ for signed-score selection, and $0.0010897$ for the fixed 16,000-path candidate. All criteria fit the full candidate pool.

For signed-score selection minus dual selection, the mean paired root-level regret difference is $0.0001551$ on the original roots (95% paired $t$ interval $[-0.0000147, 0.0003249]$), $0.0000786$ on the new roots ($[-0.0000421, 0.0001993]$), and $0.0001168$ pooled ($[0.0000151, 0.0002185]$). The median root-level difference is zero in each cohort. These comparisons describe the two frozen criteria on this candidate pool; the dual's certified quantity is the stopping regret interval.

**Computation.** The primary study takes 45.03 seconds using eight CPU workers. Summed worker elapsed times are 47.28 seconds for training-path sampling, fitting, and conditional means; 29.64 seconds for validation scoring; 200.82 seconds for test scoring and exact diagnostics; and 26.48 seconds for inner-simulation evaluation and selected-score output. These are concurrent worker times, not additive wall-clock stages. Graph construction and outer-path generation lie outside those component timers. The support refinement takes a further 12.75 seconds, and exhaustive evaluation of the full frozen family takes 14.85 seconds. Both extensions retain the original fits, validation choices, and test samples.

**MiniShed training comparison.** MiniShed has three seats and twelve cards (ranks 1–6, two copies), four cards each. A move is a single or a same-rank pair; a response uses the same count and a higher rank, or passes. After two passes the table clears and its owner leads. Rank payoffs are $(+2, 0, -2)$. Exhaustive search permits true-deal, pooled and information-aware reference checks. An assignment-enumeration sure-win implementation agrees on $398/398$ audited states; pooled soundness is checked on $400$ gate states and $801$ campaign candidates.

The multilayer perceptron (MLP) comparison uses a $24$-input, $64$-unit tanh hidden layer, $13$ action logits and a value head, $30{,}000$ training episodes, policy learning rate $0.02$, auxiliary rate $0.05$ and reset fraction $0.25$. It changes the reset/loss condition at the same fixed budget and evaluates the held-out set across five paired independent runs.

| MiniShed MLP arm | Retention (mean±SE) | Change vs. plain [95% CI] |
|---|---|---|
| Plain | $0.9357 \pm 0.0078$ | — |
| Certified resets | $0.9490 \pm 0.0093$ | $+0.0133$ $[-0.0125, 0.0391]$ |
| Certified resets + aux | $0.9730 \pm 0.0044$ | $+0.0373$ $[0.0108, 0.0638]$ |
| Uncertified resets | $0.9230 \pm 0.0092$ | $-0.0127$ $[-0.0452, 0.0199]$ |

Table 21: The complete five-run MLP reset comparison. The matched auxiliary-versus-resets effect is $+0.0240$ $[0.0032, 0.0448]$.

The last-card consequence check generates $600$ small Big Two post-action states, finds $467$ exact denials, and enumerates $1{,}119$ distinct state/terminal-outcome pairs over $9{,}094$ recursion states. Every pathwise payoff and first-place comparison agrees with the exact last-card identity. Separate checks verify exact binomial lower-tail coverage on $18{,}018$ parameter cases, with zero violations. These are executable checks of theorem consequences, in addition to the proofs and engine/search parity gates.

The natural Big Two coverage probe solves $12{,}000$ own-play decisions. The $956$ eligible positions have both a true-deal winning action and a losing action; $478$ pass the pooled test. For at most nine unseen cards, the probe enumerates every hidden split and asks whether each fully observed deal is winning. This per-deal condition is necessary for a common information-feasible strategy. Every one of the $328$ positions satisfying it also has a pooled certificate on this subpopulation, so the sufficient and necessary tests coincide there. This coverage result concerns the enumerable subpopulation.

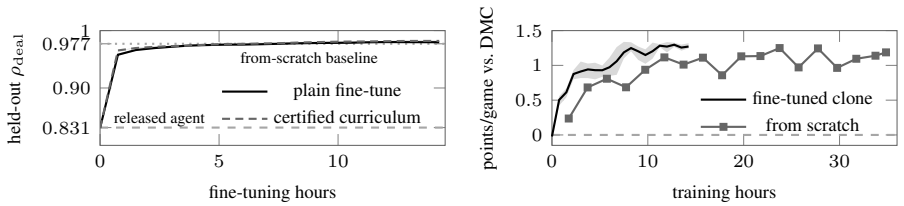



Figure 1: Big Two v2 learning curves on an RTX 5090. Left: held-out true-deal retention of the behavior-cloned released agent, fine-tuned with plain RL or certified resets. Curves show five-run means and ranges across runs. Right: logged in-training matchup scores against released DMC; the from-scratch curve is one run. Time includes in-loop evaluation.

# I PARTNERSHIP AUDIT IN GUANDAN

**Complete-round strength evaluation.** Guandan has two partnerships, two decks and 27 cards per seat. The first finisher's team receives $+3$, $+2$ or $+1$ points when its partner finishes second, third or fourth; the other team receives the negative score. We evaluate single rounds with independently sampled level and first actor, without tribute or multiround progression. Each candidate controls both seats of one partnership and samples its native categorical policy. The released DanZero DMC (Lu et al., 2022) greedily scores its full physical-card legal action list.

We freeze the final plain self-play and puzzle-trained checkpoints, each after $50.0$ billion local training decisions. The puzzle run starts from a saved plain-policy checkpoint and then uses its own $50.0$-billion-decision schedule; it injects puzzle states into $10\%$ of redealt lanes. These are two trained endpoints with distinct training histories. The policy has six residual trunk layers of width $768$, feed-forward width $1{,}536$, a $376$-action categorical head and a privileged value head. Its actor excludes the private opponent-card block. The deployed actor uses its original abstract action materialization; DanZero chooses physical cards directly.

| Final policy | Points/round | Game SE | $95\%$ interval | Win rate |
|---|---|---|---|---|
| Plain self-play | $+0.26647$ | $0.00954$ | $[0.24777, 0.28516]$ | $51.195\%$ |
| Puzzle warm start | $+0.46448$ | $0.00941$ | $[0.44603, 0.48293]$ | $54.735\%$ |

Table 22: Each frozen endpoint against released DanZero DMC on $60{,}000$ complete matched rounds, with $30{,}000$ rounds per candidate partnership parity. Intervals are mean $\pm 1.96$ game SE. Each assigned round completes; no game is excluded.

The paired endpoint difference is $+0.19802$ $[0.18079, 0.21524]$ points per round. Finite-sample one-sided empirical Bernstein lower bounds are $0.23970$ and $0.43806$ for the two endpoints; the independent, potentially nonidentical version of Maurer & Pontil (2009, Theorem 11) accommodates fixed parity strata. These strength measurements do not require attributing the difference to a single training component. Collection takes $29.80$ and $44.10$ seconds, including physical-move assertions and state capture on a shared RTX 5090.

**A certificate measurable by the opposing partnership.** Let $V$ be the audited partnership, $C = V^c$ the responding partnership, and $U$ their signed team payoff. Members of $C$ share their two hands and public history. This determines the union of $V$'s cards but not their allocation between its two seats. We retain both counts and let each victim seat play any legal combination from that union within its count. Both coalition seats are cooperative choice nodes; both victim seats are universal-response nodes. A coalition strategy that forces the first finish in this relaxed game guarantees $U \leq -1$ across every consistent hidden allocation.

For an exact physical-action quotient, the audited domain has no prior finisher, one to four cards in every hand and no remaining level-heart wild cards. Its possible new plays are singles, pairs, triples, four-card bombs and the four-joker bomb. Suits cannot affect future legality in this domain, so rank-count actions represent every physical materialization. The inherited table type and comparison key remain explicit, including larger combinations played earlier. The domain predicate and every coalition search input are measurable from its joint information.

Soundness follows by a simulation relation. Map each real victim allocation to its union and two counts. Every real victim play appears in the relaxed action set, and updating its cards and count preserves the relation. Every coalition play removes cards from its known acting hand. A positive search result supplies one choice at each observable coalition state, shared across all compatible victim allocations. Following these choices defeats every real victim continuation until the coalition's first finish. The payoff rule then supplies the cap. Search has no depth limit; exhausting its $250{,}000$-node allowance yields unknown. Failure to certify the relaxed game is not a terminal-loss label for the actual partnership.

**Natural-state evaluation.** A separate $20{,}000$-round sample uses the final puzzle policy on all four seats. We query the first post-victim-action state satisfying the domain predicate in each round and switch only when that state is certified. This selects $1{,}428$ queries before inspecting outcomes. Search certifies $644$, with zero unknowns. In ordinary play, the audited team nevertheless wins four of these $644$ rounds. Independent exhaustive physical-action replay verifies all $644$ witnesses on their recorded allocations, visiting $5{,}423$ states and $5{,}765$ edges with no escape. Search takes $0.168$ seconds; search, witness verification and analysis together take $2.82$ seconds after loading states. Ordinary collection takes $3.06$ seconds.

The complete-game cap score is $\mathbf{1}\{\text{certified}\}(U+1)$, including zero for every untriggered round. Its mean is $-0.03475$; the finite $95\%$ interval is $[-0.04505, -0.02065]$. This first-boundary cap score includes $305$ ordinary $-3$ outcomes and $95$ ordinary $-2$ outcomes at triggers. It measures the specified conservative boundary rule; certificate validity and actual response payoff are separate quantities.

We also execute all $644$ refutations against fresh sampled victim actions and complete every round. The coalition follows its observable witness until its first finish, then resumes ordinary play. Victim payoffs are $-3$, $-2$ and $-1$ in $274$, $94$ and $276$ cases; every execution respects the cap, with no illegal action, fallback or unfinished round. Across all $20{,}000$ ordinary games, actual signed gain is $-0.00265$, with a finite stratified $95\%$ interval $[-0.00776, 0.01428]$. It exceeds the cap score on every path where the two differ. Strategy preparation takes $0.391$ seconds and full-round execution takes $1.114$ seconds for $8{,}309$ decisions. An exact repeat verifies every action, physical move and terminal reward.

The cap score has support $\{-2,-1,0,2,3,4\}$; actual gain has support $\{-2,-1,0,1,2,3,4,5,6\}$. Each parity stratum contains $10{,}000$ independent rounds. For a score with $m$ possible nonzero atoms, we give each atom's two Clopper–Pearson probability tails in each of the two strata error budget $0.05/(4m)$, including atoms with zero observations. Thus each tail receives $0.0025$ for the cap and $0.0015625$ for actual gain. To form the lower expectation endpoint, multiply positive atoms by their lower probability limits and negative atoms by their upper limits; reverse these choices for the upper endpoint, then average the two strata. The union bound gives each reported interval marginal coverage at least $95\%$.

Implementation checks match the reference on $864$ physical moves from $500$ independent states. Minimax covers $96$ roots and all $495$ hidden partitions; the $45$ positive roots pass $184$ partition-specific witness checks over $1{,}236$ adversarial physical edges. The DanZero bridge passes $2{,}880$ full-precision action-value checks and $38$ exact unequal-quota round replays. All $140{,}000$ assigned Guandan rounds finish legally.